Comprehensive Psychiatry

Evaluation of whether commonly used risk assessment tools are applicable to women in forensic psychiatric institutions

--Manuscript Draft--

Manuscript Number:	COMPRPSYCHIATRY-D-24-00352R1
Article Type:	Research Paper
Keywords:	violent recidivism, recidivism, female offenders, mentally ill offenders, risk prediction, recidivism prognosis
Abstract:	Objective: By providing a structured assessment of specific risk factors, risk assessment tools allow statements to be made about the likelihood of future recidivism in people who have committed a crime. These tools were originally developed for and primarily tested in men and are mainly based on the usual criminological background of men. Despite significant progress in the last decade, there is still a lack of empirical research on female offenders, especially female forensic psychiatric inpatients. To improve prognosis in female offenders, we performed a retrospective study to compare the predictive quality of the following risk assessment tools: PCL-R, LSI-R, HCR-20 v3, FAM, and VRAG-R. Method: Data were collected from the information available in the medical files of 525 female patients who had been discharged between 2001 and 2017. We examined the ability of the tools to predict general and violent recidivism by comparing the predictions with information from the Federal Central Criminal Register. Results: Overall, the prediction instruments had moderate to good predictive performance, and the study confirmed their general applicability to female forensic psychiatric patients. Conclusion: The LSI-R proved to be particularly valid for general recidivism, and both, LSI-R and HCR-20 v3, for violent recidivism.

Ms. Ref. No.: COMPRPSYCHIATRY-D-24-00352
Comprehensive Psychiatry
Title: Evaluation of whether commonly used risk assessment tools are applicable to women in forensic psychiatric institutions

Dear Reviewers, we appreciate the time and effort you have dedicated to providing your valuable feedback on our manuscript, and we are grateful for your constructive, insightful, and encouraging comments. Please find below our replies on an item-by-item basis (in italics).

Reviewer #1:
Comments to the Author: The authors have specifically looked at cut off scores and in several places in the article, the authors recommend adjusting cut-off values for higher sensitivity. For research purposes, this makes sense, however, the HCR-20 authors emphasize that the tool is an SPJ tool, and they do not recommend using cut-off scores when used in daily practice as far as I know. So, with respect to the recommendation, I do not think this would be a good advice for practitioners with respect to the HCR-20 V3, as this is an SPJ tool, and the items should not be added up, but interpreted according to SPJ method. This is in contradiction with the advice in the manual / from the developers. Please be clear about this throughout the text.

Authors' reply: *Thank you very much for this important suggestion. We agree with the reviewer and have removed the HCR-20 v3 and FAM cut-off values from the results. Instead, we now additionally report means and standard deviations for the groups of non-recidivist and recidivist patients (see Tables 3-6).*

Comments to the Author: Introduction:
When summarizing the literature, it is sometimes not clear if the authors are referring to the larger population of justice-involved women or a more specific group: mentally ill justice-involved women. This may be important for the results of this study (as also described later in the article). In general, risk assessment tools perform better in the larger population, and less in (severely) mentally ill women.

Authors' reply: *The respective reference group has been made clearer (see page 2 and 3).*

Comments to the Author: Participants:
I would like to know some more details about the participants to be able to put the results into perspective (also given the remark above and to be able to compare to previous studies). With respect to the main clinical diagnoses: what about comorbidity? For example, how many women with substance-related disorders also have a personality disorder? I would also be interested to know a bit more about the treatment these women have received. Is it possible to include a brief paragraph about the setting?

Authors' reply: *The frequencies of comorbid personality disorders have been added to Table 2, as follows:*

Main clinical diagnosis and comorbid personality disorder[bc]	
F0: Organic disorder	4 (1%)
F10: Alcohol-related disorder	50 (10%)

F10 + F6: Alcohol-related disorder and personality disorder	21 (4%)
F11-18: Substance-related disorder to specific substance	118 (23%)
F11-18 + F6: Substance-related disorder to specific substance and personality disorder	10 (2%)
F19: Multiple drug use	157 (30%)
F19 + F6: Multiple drug use and personality disorder	37 (7%)
F2: Schizophrenic disorder	73 (14%)
F2 + F6: Schizophrenic disorder and personality disorder	8 (2%)
F3: Mood disorder	4 (.7%)
F3 + F6: Mood disorder and personality disorder	1 (.2%)
F4: Adjustment disorder / PTSD	1 (.2%)
F4 + F6: Adjustment disorder/PTSD and personality disorder	1 (.2%)
F6: Personality disorder	33 (6%)
F7: Mental retardation	1 (.2%)
F9: Conduct disorder	1 (.2%)
F9: Conduct disorder and personality disorder	2 (.4%)

We have added a short paragraph about the setting and treatment to the sample description, as follows:
Page 4: "Patients admitted under Section 63 or 64 are treated in specialized secure hospitals, where they are cared for by doctors, psychologists, and nurses rather than being supervised by security personnel. Currently, forensic psychiatric treatment focuses on addressing individual risk factors, such as specific symptoms and behaviors related to the offense, with the aim to minimize the risk of reoffending."

Comments to the Author: Can the authors briefly explain what is considered a serious crime in the German system: e.g., does this always involve violence?

Authors' reply: *We have added the definition of a serious crime, as follows:*
Page 4: "According to the German criminal code, a crime or recidivism is serious when the victims of the offence experience or are exposed to a considerable danger of severe emotional trauma or physical injury or the crime or recidivism causes serious economic damage."

Comments to the Author: 32 were excluded because they died or because of other criteria. How many died exactly?
Do the authors know more about the deceased women (age, cause of death)? See also the high mortality rate De Vogel et al. (2019) found.

Authors' reply: *Of the 32 female patients excluded from the study, 13 had passed away. Unfortunately, no information is available on the age at death or the cause of death.*
We have made the following changes to the manuscript:
Page 4: "Of these patients, 32 were excluded from further analysis because they had died (n = 13) or did not meet the inclusion criteria (n = 19)."

Comments to the Author: Instruments:
HCR-20 V3: I would advise the authors to explain more about the SPJ method and the reason why they decided to use scores (0,1,2) for research purposes instead of codes (no, partially, yes), as this may be confusing for practitioners. Later, in the Discussion, they briefly mention this, but this should already be done in the Method section. Same goes for the adaptations to the LSI.

Authors' reply:
The different use of the assessment tools LSI-R and HCR-20 v3 is thoroughly explained in the Methods and Discussion section of the revised manuscript.

Methods, Risk assessments (page 5): "In criminal prognosis, a distinction is made between actuarial risk assessment and the structured professional judgment approach: The former provides standardized and quantitative risk predictions based on statistical models (e.g., LSI-R, VRAG-R), whereas the latter offers a more flexible and nuanced approach that incorporates clinical expertise and case-specific information for assessing and managing risks (e.g., HCR-20 v3, FAM)."

Methods, LSI-R (page 7): "Two issues related to the LSI-R must be mentioned here: First, the LSI-R is intended to be completed during an interview, which was not possible in the context of the present study, and second, some of the LSI-R items were difficult to apply to the conditions of a forensic psychiatric facility, so we had to adapt them for the present study."

Methods, HCR-20 v3 (page 7): "The use of a structured professional judgement approach means that the scale is not actually designed to quantify items, including summing the fulfilled risk factors. Instead, it relies on the professional's experience and subjective interpretation of the data. According to the HCR-20 v3 manual, the professional judgment consists of seven steps: gathering case information, assessing risk factors, assessing the relevance of risk factors, risk conceptualization, risk scenarios, risk management strategies, and final judgment. In the present study, the risk assessment was based on records, so we were not able to carry out steps 3 to 7. In accordance with Brookstein [37], we assessed the presence of risk factors (step 2) by using a 3-point scale (present = 2, partially present = 1, not present = 0) and summed the scores of the 20 risk factors, which yielded a total score ranging from 0 to 40."

Discussion (page 22): "Second, for the same reason, we were not able to apply the PCL-R and LSI-R in interview form. Third, the HCR-20 v3 is not designed for quantifying items, which limits the transferability of our results; however, the AUC value for predicting a violent offense based on the HCR-20 v3 total score was very good compared with that of the other instruments, suggesting that summing the risk factors yields excellent results, obviating the need to implement steps 3 to 7. Therefore, in routine clinical care, if professionals can only assess a patient based on records, they can achieve a good prognosis by summing the fulfilled factor values."

Comments to the Author: Did the raters also code the Final Risk Judgements?

Authors' reply: *Yes, the five research assistants (who were trained in using the prognostic instruments and who independently assessed 11 patients with all five instruments to confirm*

a uniform standard of assessment ratings across reviewers) also coded the final risk judgements.
Pages 8-9: "Then, the staff members assessed patients by referring to the patient medical records and coded the final risk judgements."

Comments to the Author: Dichotomizing: PCL-R > 24 I do not think this is in the PCL-R manual, the authors usually state that the tool is also valid for women

Authors' reply:
The German manual of the PCL-R published by Mokros et al. (2017) may differ from the English original. On page 40 of the manual by Mokros et al. (2017), Table 8 contains a descriptive schema for classifying PCL-R total scores and includes the following information:

PCL-R Total Score	*Degree of Expression*	*Description*
33-40	*5*	*Very high*
25-32	*4*	*High*
17-24	*3*	*Medium*
9-16	*2*	*Low*
0-8	*1*	*Very low*

We used this classification and chose the cut-off value of 24, which distinguishes between individuals with a medium and those with a high to very high psychopathy score.
We have added the following passage to the Methods section:
Methods, PCL-R (page 6): "In accordance with the descriptive schema for classifying PCL-R total scores [29, page 40], we chose the cut-off value of 24, which distinguishes between individuals with a moderate and those with a high to very high psychopathy score."

Regarding the applicability of the instrument in women, the manual by Mokros et al. (2017) contains the following passage (page 68): "To the knowledge of the authors of this German version of the PCL-R, there are only three studies from German-speaking countries in which the PCL-R has been used with female offenders (Lehmann & Ittel, 2012) or female assessment subjects (Eisenbarth, Osterheider, Nedopil & Stadtland, 2012; Hauschild, 2014). The sample sizes were 51 (Lehmann & Ittel, 2012), 80 (Eisenbarth et al., 2012), and 63 women (Hauschild, 2014). Given the discrepancy in reported sample means (11.99 by Eisenbarth et al., 2012; 5.52 by Hauschild, 2014; and 16.15 by Lehmann & Ittel, 2012) and considering the different modes of data collection (record review only by Eisenbarth et al., 2012, and Hauschild, 2014, versus interview and record review by Lehmann & Ittel, 2012), it is difficult to integrate these findings into a cohesive overall picture. Therefore, the authors of this German version advise against the current use of the PCL-R with women in German-speaking areas for practical application."

To answer your question, no, the application of the German version of the PCL-R is not yet recommended for women because of the small and heterogeneous samples. This is precisely where the strength of the present study lies because it represents the largest German female sample to date (N = 525).

Comments to the Author: Procedure:
Can the authors explain a bit more about these research staff members: what is their background (psychologists?), and do they have any clinical experience?

Authors' reply: *Each of the five research staff members has a Master's degree in psychology and has been working as a clinical psychologist in forensic psychiatry for an average of two years. By participating in the present study, the five staff members will be able to obtain a doctorate (PhD).*
We have added the following passage to the Methods section:
Page 8: "Before the study, five research staff members (clinical psychologists) were trained in the prognostic instruments."

Comments to the Author: 11 cases were used for interrater reliability; this is a rather low number: only 2% of the sample. Was it not possible to have a larger group for interrater reliability?

Authors' reply: *The minimum number of cases required for calculating the ICC is independent of the total sample size. According to Wirtz and Caspar (2002), a minimum of 10 participants (ideally all rated by the same assessors) should be evaluated. With 11 cases (and five raters), we exceed this limit.*

Source: Wirtz, W. & Caspar, F. (2002). Beurteilerübereinstimmung und Beurteilerreliabilität. Hogrefe, Göttingen.

Comments to the Author: Were there missing data? How did the authors handle missing data?

Authors' reply: *Missing values in the sociodemographic and forensic-psychiatric variables were indicated as a note below Tables 1 and 2. In cases in which insufficient or missing information did not allow us to assess an item of 1 of the 5 assessment instruments, we followed the procedures outlined in the manuals. If essential information was missing (i.e., the presence of at least 1 written court judgment or 1 written psychiatric or psychological expert report), participants were excluded from the analysis (n = 19, see page 4).*

Comments to the Author: Outcome measure:
General recidivism: are ''documented offenses'' all formal convictions?

Authors' reply: *Yes, the documented offenses all result in a formal conviction. We have added the following phrase to the Methods section:*
Page 9: "To evaluate actual relapses after patients had been discharged, we obtained information from the German Federal Central Criminal Register in September 2020 and February 2021, in which all formal convictions are documented. Each formal conviction documented after release from the hospital or prison (in the case of treatment discontinuation) was counted as recidivism."

Comments to the Author: Statistical analyses: the authors mention to look at group differences. Was their goal to examine group differences? Which groups?

Authors' reply: *We apologize for this oversight. Originally, we had intended to statistically compare recidivists and non-recidivists, but then we removed this analysis from the manuscript. This analysis has been reinserted into the current version of the manuscript, and we have made slight revision the paragraph.*

Comments to the Author: Results:
The authors correctly state that the FAM is not a stand-alone test and they have calculated the prognostic performance of the two tools combined. However, in the Tables, it seems that the FAM is being reported as stand alone. I would recommend to also report the FAM in combination with the HCR-20 V3 in the Tables.

Authors' reply: *The AUC values for the combined assessment of HCR-20 v3 and FAM have been included in Tables 3 and 5. Sensitivity, specificity, PPV, and NPV could not be calculated because—in accordance with the reviewers' recommendations—the revised version of the manuscript no longer includes cut-off values for HCR-20 v3 and FAM.*

Comments to the Author: Was it possible to also look at predictive validity of the individual items of the tools? There are probably too many to report on in this article but could be interesting as supplemental material.

Authors' reply: *The AUC values and correlation coefficients of the individual items are reported in Supplement 1.*

Comments to the Author: DISCUSSION
The authors state: The present study aimed to improve prognosis of recidivism in women treated in forensic psychiatric facilities by evaluating the predictive quality of common prognostic instruments.
It seems more like their goal was to examine prognosis instead of improve prognosis, like stated in the Introduction:
The present study aimed to examine various risk assessment tools and identify the most appropriate one in terms of the applicability to mentally ill female offenders

Authors' reply: *You are absolutely right, and we have revised the sentence accordingly.*

Comments to the Author: The comparison with male samples including the tables is interesting, but I do not think this fits in the Discussion part. It seems to be too detailed for in the Discussion part. Also, there are new findings reported here. I think it is more logical to describe this part as post hoc analyses, or possibly in supplemental material and in the Discussion refer to it in a more descriptive way.

Authors' reply: *The comparison with the male sample has been moved to Supplement 2.*

Comments to the Author: The authors mention here that they have made some adaptations to the LSI: this should be mentioned already in the Method section.

Authors' reply: *The adaptations to the LSI-R are now also described in the Methods section (see pages 6-7).*

Comments to the Author: The clinical implications remain somewhat superficial: can the authors be more specific, for instance, about adjusting cut off values (and only for the tools that are to be used in an actuarial way). What are practical advises for practitioners performing risk assessments? What could, for instance, be the clinical value of the tools next to prognosis?

Authors' reply: *The Discussion, including the clinical recommendations, has been thoroughly revised.*

Comments to the Author: Do the authors have suggestions for future research?
Authors' reply: *We added the following passage to the Discussion section: "Future studies could examine various measures to enhance the sensitivity of instruments used in forensic psychiatric samples of women, such as altered cut-off scores, different weighting of individual risk factors, and the combined application of different instruments."*

Reviewer #2:
Comments to the Author: p.7. I was bit surprised the authors recommended a cut-off of the PCL-R, HCR-20, LSI, VRAG-R and FAM from their specific manuals, could they indicate the pages of these recommendations?

Authors' reply: *We apologize that we did not clearly explain how the cut-off values were defined. Please see below:*

PCL-R: On page 40 of the manual by Mokros et al. (2017), Table 8 contains a descriptive schema for classifying PCL-R total scores and includes the following information:

PCL-R Total Score	*Degree of Expression*	*Description*
33-40	*5*	*Very high*
25-32	*4*	*High*
17-24	*3*	*Medium*
9-16	*2*	*Low*
0-8	*1*	*Very low*

We used this classification and chose the cut-off value of 24, which distinguishes between individuals with a moderate and those with a high to very high psychopathy score.
We have added the following passage to the Methods section of the revised manuscript:
Page 6: "In accordance with the descriptive schema for classifying PCL-R total scores [29, page 40], we chose the cut-off value of 24, which distinguishes between individuals with a moderate and those with a high to very high psychopathy score."

LSI-R: Page 72 of the manual by Dahle et al. (2012) contains a descriptive schema for classifying LSI-R total scores and includes the following information:

LSI-R Total Score	*Description*
> 40	*High risk*
34-40	*Increased risk*
24-33	*Moderate risk*
14-23	*Low to moderate risk*
0-13	*Low risk*

We used this classification and chose the cut-off value of 33, which distinguishes between individuals with a moderate and those with an increased or high risk.
We have added the following passage to the Methods section of the revised manuscript:

Pages 6-7: "In accordance with the LSI-R recidivism risk classification [34, page 72], we chose the cut-off value of 33, which distinguishes between individuals with a moderate and those with an increased or high risk."

VRAG-R: On page 5 of the manual by Rettenberger et al. (2017), you will find a table with 9 risk categories of the VRAG-R that includes the following information:

Risk category	*VRAG-R Total score*	*Recidivism probability after 5 years*	*Recidivism probability after 12 years*
1	*≤ -24*	*9%*	*15%*
2	*-23 to -17*	*12%*	*24%*
3	*-16 to -11*	*16%*	*33%*
4	*-10 to -4*	*20%*	*42%*
5	*-3 to +3*	*26%*	*51%*
6	*+4 to +11*	*34%*	*60%*
7	*+12 to +17*	*45%*	*69%*
8	*+18 to +26*	*58%*	*78%*
9	*≥ +27*	*76%*	*87%*

We used this classification and chose the cut-off value of 11, which distinguishes between individuals with recidivism rates < 45% and ≥ 45% (after 5 years) and < 69% and ≥ 69% (after 12 years).

We have added the following passage to the Methods section of the revised manuscript Page 8: "In accordance with the risk categories of the VRAG-R [38, page 5], we chose the cut-off value of 11, which distinguishes between individuals with recidivism rates below 45% and equal to or greater than 45% after 5 years and below 69% and equal to or greater than 69% after 12 years."

HCR-20 v3 and FAM: Because HCR-20 v3 and FAM use the structured professional judgment (SPJ) approach, Reviewer 1 suggested that we refrain from using cut-off values for these methods.

Comments to the Author: Confidence Intervals should be mentioned for each instrument in all tables.

Authors' reply: *We have added confidence intervals to the AUC values (see Tables 3 and 5).*

Comments to the Author: p. 15 Youden index should be further described in a full sentence than a (sensitivity, 1-specificity).

Authors' reply: *In the revised manuscript, we have deleted the calculation of the Youden index.*

Comments to the Author: The conclusion should discuss further the fact that VRAG-R was much more predictive than the FAM (Table 4).

Authors' reply: *According to the authors of the FAM, the instrument was developed as an extension of the HCR-20 v3, and it is not intended to be used in isolation. Therefore, in the revised version of the manuscript, we have omitted the description of its isolated use. When the FAM is used in combination with the HCR-20 v3, the two procedures perform better than the VRAG-R. We have revised the conclusions accordingly.*

Comments to the Author: Although the lower predictivity of the FAM was mentioned, more is needed in term of recommendation concerning this last instrument. Hence, I would like to hear more on the recommendations of the authors in term of prioritization and complementarity of each instrument with others for the assessment of female offenders.

Authors' reply: *In the revised version of the manuscript, the predictive performance of the individual instruments is also statistically compared. Here, the combined application of FAM and HCR-20 v3 was found to be less effective than the use of HCR-20 v3 alone. We mention this in the Discussion and discourage additional use of FAM. In addition, we have thoroughly revised the Discussion, including the clinical recommendations.*

#Reviewer 3

Comments to the Author: 1. Introduction

Page 2, Line 2: It is essential to expand the first paragraph concerning mental illness in criminal justice populations to better align with the scope of this journal. While you elaborated on the issues of predictive validity of risk assessment tools among women with a history of crimes in the second paragraph, the first paragraph concerning mental illness and recidivism is relatively brief. I suggest adding more relevant literature about the relationship between targeting mental health and recidivism, as well as the prediction of risk for mentally disordered offenders.
Here are some examples of relevant literature to consider:

Ogilvie, J.M., Tzoumakis, S., Thompson, C. et al. Psychiatric illness and the risk of reoffending: recurrent event analysis for an Australian birth cohort. BMC Psychiatry 23, 355 (2023). https://doi.org/10.1186/s12888-023-04839-0

Ogonah, M. G. T., Seyedsalehi, A., Whiting, D., & Fazel, S. (2023). Violence risk assessment instruments in forensic psychiatric populations: a systematic review and meta-analysis. The lancet. Psychiatry, 10(10), 780–789. https://doi.org/10.1016/S2215-0366(23)00256-0

Okamura, M., Okada, T., & Okumura, Y. (2023). Recidivism among prisoners with severe mental disorders. Heliyon, 9(6), e17007. https://doi.org/10.1016/j.heliyon.2023.e17007

Zgoba, K. M., Reeves, R., Tamburello, A., & Debilio, L. (2020). Criminal recidivism in inmates with mental illness and substance use disorders. Journal of the American Academy of Psychiatry and the Law, 48(2), 209–215.
By incorporating these studies and others, you can provide a more comprehensive background that aligns with the journal's focus on understanding mental illness and psychopathology.

Author's reply: *Thank you very much for the references. We have added the following to expand on the first paragraph about mental illness in criminal justice populations:*

Page 2: "Research by Ogilvie et al. [6] found that a higher percentage of mentally ill individuals than individuals without psychiatric diagnoses returned to court for all categories of offenses, including violent (20.5% vs 8.5%, respectively), nonviolent (60.3% vs 40.3%, respectively), and other minor offenses (61.7% vs 44.1%, respectively). Co-occurring substance use disorders further exacerbate recidivism rates [7]. Furthermore, Okamura et al. [8] demonstrated that individuals who recidivate often lack medical services and support."

In addition, on page 3 we have added more information on the prediction of risk in mentally disordered offenders: "Recently, Ogonah et al. [12] conducted a meta-analysis to evaluate the performance of these tools in forensic mental health contexts. They reported pooled areas under the curve (AUCs) ranging from .64 to .74, predominantly in male samples (only two out of 50 studies assessed risk assessment tools in female-only samples)."

Comments to the Author: Page 3 Line 5: I suggest adding more discussion about mental illness issues beyond just the female population. This would provide a broader context and highlight the general challenges and considerations in assessing risk among individuals with mental health disorders.

Authors' reply: *We have added more information about mentally ill offenders on page 2 because it seemed more appropriate to address it there, given that page 3 already explicitly discusses the topic of women:*
Page 2: "Mental illness is prevalent among criminal justice populations, and studies indicate that approximately 10% to 15% of inmates have psychotic disorders, 20% to 30% experience depression and bipolar disorders, and over 50% have a substance use disorder [1,2]. Factors contributing to the relationship between mental disorders and criminal behavior include emotion dysregulation as a core feature of many psychiatric disorders; cognitive distortions, such as irrational beliefs or misinterpretations of social cues; and difficulties in regulating behavior, including impulsivity and poor impulse control [3–5]. Mental disorders increase not only the risk of committing a crime but also the risk of recidivism."

Comments to the Author: 2. Material and Methods, Page 4 Line 58: Please note that the PCL-R was not originally designed as a risk assessment tool, but it is commonly used because of its ability to predict recidivism.

Authors' reply: *We have added the following sentence on page 5 of the revised manuscript: "It was not originally designed as a risk assessment tool, but it is commonly used as a prognostic tool because of its ability to predict recidivism."*

Comments to the Author: Page 4 Line 29: It would be helpful to define "prognosis" versus "diagnosis" as commonly used in medical evaluation. Prognosis refers to the likely course and outcome of a condition, while diagnosis involves identifying the nature of an illness or problem through examination of the symptoms.

Authors' reply: *The definition has been added (see page 5).*

Comments to the Author: Page 5 Line 10: *r* values were reported for predictive validity under section 2.2. However, Author(s) used AUC values to evaluate the predictive validity with current samples. It would be helpful to provide more information on the differences, pros, and

cons of these different metrics (at least r and AUC). Explain why author(s) selected AUC over *r*, and provide benchmarks for *r* values to help readers understand and interpret these values.

Authors' reply: *In the context of prognosis, both r (often point-biserial correlation values) and AUC (Area Under the Receiver Operating Characteristic Curve) are used as metrics to assess the performance or predictive ability of tools or variables; however, they measure different aspects: Correlation analysis is a statistical procedure for measuring and evaluating the strength and direction of the relationship between two variables, and it allows one to determine whether changes in one variable are associated with changes in another variable and quantifies the degree of this connection; AUC is used to evaluate the overall discriminatory ability of a prognostic tool and indicates how well the tool can distinguish between different outcomes (e.g., individuals who relapse vs those who do not). Correlation analysis is often used to evaluate the strength of a linear association between a single prognostic factor and an outcome, whereas AUC is used to assess the overall predictive performance of a model.*

In the Results section of the revised manuscript, we now provide r values in addition to AUCs (see Tables 3 and 5). Both metrics are introduced in the Methods section, and we also give benchmarks for r:

See page 10: "Point-biserial correlation analysis is a statistical procedure for measuring and evaluating the strength and direction (i.e., positive versus negative correlation) of the relationship between two variables. It allows one to determine whether changes in one variable are associated with changes in another and quantifies the degree of this connection (strong effect, ± .50 and above; medium effect, between ± .30 and ± .49; small effect, .29 and below [40])."

Comments to the Author: Page 5 Line 41: As mentioned, the HCR-20 is a structured professional judgment tool. It would be worthwhile to mention how this approach (professional judgment after summing up the scores) differs from other approaches (e.g., actuarial) and how you applied this approach. This earlier explanation would help readers understand the limitation stated on page 16, line 24.

Authors' reply:
The different approach of the HCR-20 v3 is thoroughly explained in the Methods and Discussion section of the revised version of the manuscript:

Methods (page 7): "The use of a structured professional judgement approach means that the scale is not actually designed to quantify items, including summing the fulfilled risk factors. Instead, it relies on the professional's experience and subjective interpretation of the data. According to the HCR-20 v3 manual, the professional judgment consists of seven steps: gathering case information, assessing risk factors, assessing the relevance of risk factors, risk conceptualization, risk scenarios, risk management strategies, and final judgment. In the present study, the risk assessment was based on records, so we were not able to carry out steps 3 to 7. In accordance with Brookstein [37], we assessed the presence of risk factors (step 2) by using a 3-point scale (present = 2, partially present = 1, not present = 0) and summed the scores of the 20 risk factors, which yielded a total score ranging from 0 to 40."

Discussion (page 22): "Third, the HCR-20 v3 is not designed for quantifying items, which limits the transferability of our results; however, the AUC value for predicting a violent offense based on the HCR-20 v3 total score was very good compared with that of the other instruments, suggesting that summing the risk factors yields excellent results, obviating the need to implement steps 3 to 7. Therefore, in routine clinical care, if professionals can only assess a patient based on records, they can achieve a good prognosis by summing the fulfilled factor values."

Comments to the Author: Page 6 Line 14: Please clarify how the original five-point scale was merged into a three-point scale. Justify this modification (i.e., to ensure better comparability within the present study). Did the user guide (manual) provide a rule to combine the FAM and HCR-20 v3? If so, briefly describe it, as this combination was used in the later analyses.

Authors' reply: *The FAM manual does not provide a rule for numerically combining FAM and HCR-20 v3. Because both approaches follow the structured professional judgment approach, neither the FAM manual nor the HCR-20 v3 manual provides instructions on how to proceed with the risk factors if one wishes to calculate a numerical risk. This topic is mentioned in the Methods section of the revised manuscript, as follows:*
Page 8: "Like the HCR-20 v3, the FAM also follows the structured professional judgement approach, meaning that ratings are not intended to be given numerical values. However, to ensure better comparability of findings in the present study, we rated the items by using an approach similar to that used for the HCR-20 v3, i.e., on a three-point scale (0 = no, 1 = possible or partial, 2 = yes), and summed the scores of the 10 FAM risk factors, which yielded a total score ranging from 0 to 20 (...) Because the combined evaluation of the HCR-20 v3 and FAM included 18 items from the HCR-20 v3 and 20 items from the FAM, the total scores ranged from 0 to 56."

Comments to the Author: Page 6 Line 46: Please clarify how many reviewers rated each patient (likely two?). It would also be beneficial if you could report the inter-rater reliability (ICCs) for each tool.

Authors' reply: *To confirm that the standard of assessment ratings was uniform across reviewers, all five reviewers independently rated 11 patients with the five assessment tools. In the revised manuscript, we report the ICCs for each tool.*
Page 9: "To confirm a uniform standard of assessment ratings across reviewers, all five reviewers independently rated 11 patients with the five assessment tools and interrater reliabilities were calculated (the ICCs of the assessment tools were as follows: PCL-R, .71; LSI-R, .74; HCR-20 v3, .61; FAM, .82; and VRAG-R, .89)."

Comments to the Author: Page 6 Line 56: Please provide a clear definition of violent and general recidivism used in this manuscript. It seems that general recidivism was defined as any offense that occurred (arrest or report?), and violent recidivism was defined as convictions.

Authors' reply: *No, only the offenses for which the patients were convicted were counted as recidivism. We have revised the Methods section accordingly:*

Page 9: "To evaluate actual relapses after patients had been discharged, we obtained information from the German Federal Central Criminal Register in September 2020 and

February 2021, in which all formal convictions are documented. Each formal conviction documented after release from the hospital or prison (in the case of treatment discontinuation) was counted as recidivism. In addition, violent offenses, which were defined as convictions for an offense involving crimes against persons (e.g., homicide, sex crimes, assault, threat, and robbery), were analyzed separately."

Comments to the Author: Page 7 Line 39: please consider reporting effect sizes in addition to the statistical significance tests (e.g., Cramer's *v* for Chi-squared tests). Additionally, I did not see where t-tests and Mann-Whitney U-tests were used. Please clarify.

Authors' reply: *We apologize for this oversight. Originally, we had intended to statistically compare recidivists and non-recidivists, but then we removed this analysis from the manuscript. This analysis has been reinserted into the current version of the manuscript, and we have slightly revised the paragraph. In addition, we report effect sizes.*

Comments to the Author: Page 7 Line 49: It is doubtful that the manuals provide a strict cut-off value, but rather cut-scores for risk categories. Could you provide a clearer explanation of what the cut-off value means in this context?

Authors' reply: *We apologize that we did not clearly explain how the cut-off values were defined. Please see below:*

PCL-R: On page 40 of the manual by Mokros et al. (2017), Table 8 contains a descriptive schema for classifying PCL-R total scores and includes the following information:

PCL-R Total Score	*Degree of Expression*	*Description*
33-40	*5*	*Very high*
25-32	*4*	*High*
17-24	*3*	*Medium*
9-16	*2*	*Low*
0-8	*1*	*Very low*

We used this classification and chose the cut-off value of 24, which distinguishes between individuals with a moderate and those with a high to very high psychopathy score.
We have added the following passage to the Methods section of the revised manuscript:
Page 6: "In accordance with the descriptive schema for classifying PCL-R total scores [29, page 40], we chose the cut-off value of 24, which distinguishes between individuals with a moderate and those with a high to very high psychopathy score."

LSI-R: Page 72 of the manual by Dahle et al. (2012) contains a descriptive schema for classifying LSI-R total scores and includes the following information:

LSI-R Total Score	*Description*
>40	*High risk*
34-40	*Increased risk*
24-33	*Moderate risk*
14-23	*Low to moderate risk*
0-13	*Low risk*

We used this classification and chose the cut-off value of 33, which distinguishes between individuals with a moderate and those with an increased or high risk.
The following passage was added to the Methods section of the revised manuscript.
Pages 6-7: "In accordance with the LSI-R recidivism risk classification [34, page 72], we chose the cut-off value of 33, which distinguishes between individuals with a moderate and those with an increased or high risk."

VRAG-R: On page 5 of the manual by Rettenberger et al. (2017), you will find a table with 9 risk categories of the VRAG-R that includes the following information:

Risk category	*VRAG-R Total score*	*Recidivism probability after 5 years*	*Recidivism probability after 12 years*
1	*≤ -24*	*9%*	*15%*
2	*-23 bis -17*	*12%*	*24%*
3	*-16 bis -11*	*16%*	*33%*
4	*-10 bis -4*	*20%*	*42%*
5	*-3 bis +3*	*26%*	*51%*
6	*+4 bis +11*	*34%*	*60%*
7	*+12 bis +17*	*45%*	*69%*
8	*+18 bis +26*	*58%*	*78%*
9	*≥ +27*	*76%*	*87%*

We used this classification and chose the cut-off value of 11, which distinguishes between individuals with recidivism rates < 45% and ≥ 45% (after 5 years) and < 69% and ≥ 69% (after 12 years).

We have added the following passage was added to the Methods section of the revised manuscript:
Page 8: "In accordance with the risk categories of the VRAG-R [38, page 5], we chose the cut-off value of 11, which distinguishes between individuals with recidivism rates below 45% and equal to or greater than 45% after 5 years and below 69% and equal to or greater than 69% after 12 years."

HCR-20 v3 and FAM: Because HCR-20 v3 and FAM use the structured professional judgment approach, Reviewer 1 suggested that we refrain from using cut-off values for these methods.

Comments to the Author: 3. Results, Page 10, Table 2: Please indicate who diagnosed the mental disorders in the patient population.

Authors' reply: *The diagnoses were made by the physicians working in the forensic psychiatric hospital. This information was added to Table 2.*

Comments to the Author: Page 11, Line 28: In addition to the predictive validity test, I would encourage the authors to explore potential moderating effects of different types of diagnoses, particularly focusing on schizophrenia and alcohol/substance-related disorders, after controlling for other risk factors. Although there is no control group (individuals without a mental disorder), conducting these additional analyses would enrich our understanding of the relationship between mental disorders and recidivism. It could shed light on how specific types

of mental disorders may interact with other risk factors to influence the likelihood of recidivism among women in forensic psychiatric institutions.

Authors' reply: *In accordance with your recommendation, we compared the predictive performance of the five instruments in two subgroups: patients with schizophrenia (n = 81) and patients with a substance use disorder (n = 393). We compared the AUCs by the non-parametric Mann-Whitney U test. The results revealed no significant differences between the two diagnostic groups in terms of the predictive validity of the five assessment tools, and the AUC differences ranged from -.007 to .106. This analysis was added to the revised manuscript (see Methods [pages 10-11], Results [page 19], and Discussion [page 21]).*

Comments to the Author: Page 12, Table 3 and 4: I encourage you to run and report item-level analyses. This would greatly contribute to integrating the cumulative evidence of the predictive accuracy of risk assessment tools and factors for men and women. Additionally, please provide the 95% Confidence Intervals for the AUC values. Also, add an additional line for the combined results of HCR-20 and FAM in the tables.

Authors' reply: *The AUC values of the individual items are reported in Supplement 1.*
The 95% Confidence Intervals for the AUC values and the combined results of HCR-20 and FAM were added to Tables 3 and 5.

Comments to the Author: 4. Discussion, Page 13, Line 15: The aim of the current study was not to improve prognosis, but rather to understand the current status of the predictive validity of risk assessment tools for women. Please rewrite this section accordingly.

Authors' reply: *The first sentence of the Discussion was revised, as follows: "The present study aimed to examine prognosis of recidivism in women treated in forensic psychiatric facilities by evaluating the predictive quality of common prognostic instruments (PCL-R, LSI-R, HCR-20 v3, HCR-20 v3 + FAM, and VRAG-R)." In addition, the Discussion has been extensively revised.*

Comments to the Author: Page 13, Line 39: Please add effect sizes for the chi-squared tests or use other effect size metrics, such as 2x2 odds ratios.

Authors' reply: *Cramer V was added. Another reviewer recommended moving the comparison between male and female patients to a supplement. Therefore, you will now find the effect sizes in Supplement 2.*

Comments to the Author: Page 13 Line 42: This study found relatively high recidivism rates among women offenders compared to findings from other literature, although the differences were not statistically significant. Could you elaborate on the potential reasons for this trend specifically for the current samples (e.g., any unique characteristics of the current sample)? Additionally, it would be beneficial to have a control group (male offenders in the institution), though this might require substantial additional resources and may not be feasible. If this is not possible, please consider this as another limitation of the current study.

Authors' reply: *The focus of the present study was to examine whether predictive instruments can be applied to women. To investigate this topic, a male control group is not*

necessary, and the comparison of recidivism rates with a male sample was just an add-on. We have now moved this add-on to Supplement 2.

The relatively high recidivism rates are discussed in Supplement 2: "Relapse rate can be influenced by many factors. For example, the length of the time at risk is important in that the longer the time at risk, the higher the relapse rate. In addition, relapse rate can also depend on the composition of the sample, and the literature provides evidence that patients with a substance use disorder are particularly prone to relapse. Both these factors are relevant in the present sample. Therefore, for our comparison with male patients, we chose a sample that had a similar time at risk and a similar prevalence of diagnoses as the female patients in our study."

By the way, it would not be possible to collect a male sample at the same hospital because the forensic psychiatric hospital in Taufkirchen treats only women.

Comments to the Author: Page 15 Line 10: What risk assessment tools were used in the Ramesh et al. (2018) study? Please list these tools. Also, in Table 5, please add AUC values.

Authors' reply: *Upon the recommendation of another reviewer, the comparison with Ramesh et al. was removed from the manuscript.*

Comments to the Author: Regarding the proposed new cut-off scores based on sensitivity and specificity: These values trade off against each other depending on samples and base rates, in addition to prediction accuracy. If you wish to include these analyses, please provide some background on this issue in the introduction, clear analytic plans, and the implications of these new cut-off scores. The current version of the manuscript lacks clear practical implications of these new scores. Also, consider moving Tables 5 and 6 to the results section. In the discussion, provide more implications of these results.

Authors' reply: *Yes, in the revised manuscript the proposed new cut-off scores have been removed.*

Comments to the Author: Page 16, Line 20: The LSI-R was originally designed for use with probationers and parolees, although it has proven useful with other community corrections samples, and within prisons, jails, halfway houses, and forensic mental health clinics and hospitals. Please clarify this point.

Authors' reply: *Thank you for this suggestion. We have added this information to the Methods section of the LSI-R (see page 6).*

Comments to the Author: In the limitations section: Was there any limitation regarding the source of recidivism information? Does the German Federal Central Criminal Register cover all regions of Germany? Please indicate any limitations related to the recidivism information, as this can have an important impact on the results.

Authors' reply: *Yes, you are right that there are additional limitations regarding the source of recidivism information, and we now mention them in the revised version of the manuscript (see page 22): "Fourth, recidivism was assessed from entries in the Federal Central Register, so incidents from the dark figure of crime were not captured. Additionally, in*

accordance with Section 46 of the Federal Central Register Act (BZRG), entries in the Federal Central Register are deleted after the expiration of a specified time limit, which depends on the severity of the offense. Minor offenses (e.g., fines up to 90 daily rates) are deleted after 3 years, offenses resulting in a sentence of no more than one year of imprisonment are deleted after 5 years, and more serious offenses (e.g., sentences exceeding one year of imprisonment) are deleted after 10 years."

Comments to the Author: References, Please ensure the required reference style of this journal (see below or refer to Guide for authors).
Reference style
Text: Indicate references by number(s) in square brackets in line with the text. The actual authors can be referred to, but the reference number(s) must always be given.
List: Number the references (numbers in square brackets) in the list in the order in which they appear in the text.

Authors' reply: *Thank you for noticing this oversight. We have revised the references.*

Reviewer #4:
Highlights:
For clarity's sake, please add 'in female patients' to the first two highlights

Authors' reply: *We have revised the highlights accordingly and added "in female forensic psychiatric patients."*

Comments to the Author: Concerning violent recidivism, most employed instruments performed well. In my opinion, this is a more important message than the finding that the HCR-20 v3 performed slightly better than the others.

Authors' reply: *In the revised version of the manuscript, we also statistically compare the predictive performances (i.e., AUC values) of the five instruments. These analyses revealed that both HCR-20 v3 and LSI-R significantly outperformed VRAG-R in predicting violent recidivism. The Discussion has been revised accordingly.*

Comments to the Author: Abstract:
Line 24: It is not mentioned that only total scores of the instruments were used.

Authors' reply: *In the revised version of the manuscript, the scales Factor 1 and Factor 2 of the PCL-R and the scales Historical, Clinical, and Risk Management of the HCR-20 v3 are now considered in the analysis of predictive validity.*

Comments to the Author: Line 49: In the conclusion it is stated that LSI-R and HCR-20 v3 proved 'particularly valid'. This is not in agreement with the results and needs rephrasing.

Authors' reply: *We have revised the sentence.*

Comments to the Author: 1 Introduction:
Line 29: reference 3 (German paper in German journal) is a bit odd, as there are many

papers on this subject that are easier to access. Please replace it by a well-known English paper.

Authors' reply: *We have replaced the reference with the following: Brown, J., & Singh, J. P. (2014). Forensic risk assessment: A beginner's guide. Archives of Forensic Psychology, 1(1), 49-59.*

Comments to the Author: Line 41: reference 5 is already part of the systematic review 6, and should thus be omitted.

Authors' reply. *Reference 5 was omitted.*

Comments to the Author: Line 44: too many references (8-19) for the point that females may have different pathways to criminal behavior than males.

Authors' reply: *The references have been reduced to the three essential ones.*

Comments to the Author: Page 3, line 2: For reasons of accessibility, reference 23 should be replaced by the English version of this paper published in Criminal Justice and Behavior, 2019, 46(4), 528-549.

Authors' reply: *The source has been replaced.*

Comments to the Author: 2.2 Risk assessments
First paragraph: please present numbers for references instead of authors and year of publication.

Authors' reply: *We have revised the references.*

Comments to the Author: First paragraph: you may consider omitting the part on the Women's Risk Needs Assessment, as it is not informative for the reader.

Authors' reply: *We have deleted the paragraph.*

Comments to the Author: 2.2.1 PCL-R
Why was only the total score analyzed and not the different factors of the PCL-R, as they may have different prediction accuracies?

Authors' reply: *In the revised manuscript, the subscales of the PCL-R and HCR-20 v3 have now been taken into account (see Tables 3 and 5).*

Comments to the Author: First paragraph: What do you mean by 'external assessment'?

Authors' reply: *We meant "third-party assessment." In the revised version of the manuscript, we have replaced "external" with "third-party."*

Comments to the Author: line 58: (is missing.

Authors' reply: *Thank you for noticing this oversight. "(" was added.*

Comments to the Author: line 61: as far as I know the PCL-R is not mainly used as a prognostic tool.

Authors' reply: *Yes, you are right. We have added the following phrase (see page 5): "It was not originally designed as a risk assessment tool, but it is commonly used as a prognostic tool because of its ability to predict recidivism."*

Comments to the Author: Page 5, line 7: consistency.

Authors' reply: *"consistencies" was deleted.*

Comments to the Author: Since most studies on predictive accuracy of the PCL-R were done in other languages than German, it would be useful to present figures from these studies too.

Authors' reply: *We have added values from international studies (see page 6): "The scale is reported to have good psychometric properties (Cronbach's α = .87, inter-rater reliability r = .91, and test-retest reliability r = .94, [30–32]).."*

Comments to the Author: 2.2.2. LSI-R
Replace authors and year of publication by reference numbers.

Authors' reply: *We have revised the references.*

Comments to the Author: Last sentence: risk of relapse to what?

Authors' reply: *The sentence was changed to "the risk of general recidivism."*

Comments to the Author: 2.2.3 HCR-20 v3:
Insert reference numbers.

Authors' reply: *Bolzmacher et al, 2014 was replaced by a reference number.*

Comments to the Author: Why was only total score analyzed and not professional judgement, total score of historical, clinical and risk management scales as well?

Authors' reply: *These scales were analyzed and are included in the revised manuscript.*

Comments to the Author: 2.2.4 FAM
The FAM also exists in Dutch and various other languages.

Authors' reply: *Thank you for this comment. We have revised the paragraph, as follows (see page 8): "For the present study, the English version of the FAM was used."*

Comments to the Author: 2.2.5 VRAG-R
Please explain the meaning of 'actuarial'.

Authors' reply: *In criminal prognosis, a distinction is made between actuarial instruments and the structured professional judgment approach. Actuarial risk assessment provides*

standardized and quantitative risk predictions based on statistical models, whereas structured professional judgment offers a more flexible and nuanced approach that incorporates clinical expertise and case-specific information in assessing and managing risks.

The different approach of each instrument is explained in the Methods section of the revised manuscript (see page 5).

Comments to the Author: Present reference numbers and not authors and years of publication.

Authors' reply: *We have revised the references.*

Comments to the Author: 2.3 Procedures
Line 51: .61 and .89 instead of .606 and .891

Authors' reply: *In the revised version of the manuscript, we report the ICCs separately for each tool and with only two decimal places: "ICCs of the assessment tools: PCL-R, .71; LSI-R, .74; HCR-20 v3, .61; FAM, .82; VRAG-R, .89."*

Comments to the Author: Page 7, first paragraph: it is likely that some patients were admitted to a mental health hospital, without (officially) having committed a crime. This probably affected recidivism risk. How was this problem handled?

Authors' reply: *In Germany, all patients in forensic psychiatric hospitals have committed a crime, and they are all admitted on the basis of a court decision. We did not investigate patients in psychiatric hospitals, for which your comment would be valid.*

Comments to the Author: I assume that all instruments for one patient were scored by the same scorer. The outcome of one assessment instrument, e.g. the PCL-R, is likely to affect the scoring of the following assessment, e.g. items of the historical scale of the HCR-20 v3. Could you describe the scoring procedure in more detail and discuss its potential impact on the data?

Authors' reply: *Yes, that's correct. All instruments for one patient were scored by the same scorer. This approach was also employed in another study: Coid, J., Yang, M., Ullrich, S., Zhang, T., Sizmur, S., Roberts, C., Farrington, D. P., & Rogers, R. D. (2009). Gender differences in structured risk assessment: Comparing the accuracy of five instruments. Journal of Consulting and Clinical Psychology, 77(2), 337-348.*

We have outlined the limitations of this approach, as follows (see page 22): "And last, all instruments were scored by one scorer, so it is possible that the outcome of one instrument may have influenced the outcome of another."
The scoring procedure is described in more detail, as follows (see pages 8-9): "Before the study, five research staff members (clinical psychologists) were trained in the prognostic instruments. Then, the staff members assessed patients by referring to the patient medical records and coded the final risk judgements."

Comments to the Author: 2.4 Statistical analysis
As the FAM has been developed as an extension of the HCR-20 v3, it should only be used as such, and not as a separate risk assessment instrument.

Authors' reply: *In the revised version of the manuscript, the FAM was only considered in combination with the HCR-20 v3.*

Comments to the Author: Last paragraph: You mention that AUC values > .714 are generally considered good. Mention in the results section which instruments performed good in predicting general and violent recidivism.

Authors' reply: *In the revised version of the manuscript, the AUC values are interpreted as follows:*
Page 14 (general reoffending): "The confidence intervals of the AUC values for all five instruments encompassed the threshold of .714, indicating that none exceeded this value. Therefore, the prognostic validity can be interpreted as moderate to good. When comparing the AUC values of the instruments, the LSI-R proved to be the best instrument for predicting general recidivism in the present sample (LSI-R compared to PCL-R: z = 2.956, p = .003, d = .260; to HCR-20 v3: z = 3.442, p = .001, d = .304; to HCR-20 v3 and FAM: z = 2.670, p = .008, d = .235; and to VRAG-R: z = 2.735, p = .006, d = .240). Furthermore, the prediction of general recidivism with the combined HCR-20 v3 and FAM was not better than that with the HCR-20 v3 alone (z = 1.780, p = .075, d = .156). When considering the scales of the PCL-R, Factor 2 was a significantly better predictor of general recidivism than Factor 1 (z = 4.298, p < .001, d = .382). The HCR-20 v3 risk management subscale also performed significantly better than the HCR-20 v3 clinical scale (z = 3.474, p = .001, d = .307). All other pairwise comparisons of the AUC metrics not mentioned did not differ significantly from each other (AUC values at the item level can be found in Supplement 1)."

Pages 16-17 (violent reoffending): "Again, the confidence intervals of the AUC values of all five prognostic instruments included the threshold of .714, so all instruments can be considered to be moderate to good. The comparison of the AUC metrics of the instruments showed the following differences: The VRAG-R predicted violent offenses less well than the HCR-20 v3 (z = 2.397, p = .017, d = .210) and the LSI-R (z = 2.074, p = .038, d = .182), and the HCR-20 v3 predicted violent offenses better than the HCR-20 v3 combined with the FAM (z = 2.504, p = .012, d = .220). All other pairwise comparisons of the AUC metrics between the instruments or the scales were not significant."

Comments to the Author: 2.5 Transparency ...
Codes instead of code

Authors' reply: *Thank you for noticing this oversight.*

Comments to the Author: 3 Results
3.1: Please briefly describe the group of patients, as this is important for the comparison with earlier results of similar studies.
Comments to the Author: 3.1: please shortly describe prevalent forensic psychiatric characteristics. For instance, the finding that most patients have a substance use disorder, only few have been diagnosed with a personality disorder.

Authors' reply: *To avoid duplicating the content of the tables in the text, we have added the following description: "All patients were female. The most common diagnoses were substance use disorder (n = 393) and schizophrenia (n = 81). About 15% of the patients had a comorbid personality disorder."*

Comments to the Author: 3.1: is an alcohol-related disorder not part of substance-related disorders to specific substance?

Authors' reply: *Yes, you are right. We have listed the alcohol-related disorders separately. In the revised table, we have included the ICD-10 codes to clarify which diagnoses are meant.*

Comments to the Author: 3.2 Predictive validity
As suggested above, omit analysis of the FAM alone and add HCR-20 v3 +FAM to Table 3 and 4.

Authors' reply: *In the revised version of the manuscript, the FAM was only considered in combination with the HCR-20 v3.*

Comments to the Author: You may also analyze whether the FAM significantly adds to the predictive validity of the HCR-20 v3 both concerning general and violent reoffending.

Authors' reply: *We have added this analysis (see page 14): "Furthermore, the prediction of general recidivism with the combined HCR-20 v3 and FAM was not better than that with the HCR-20 v3 alone (z = 1.780, p = .075, d = .156)."* ***and*** *page 17: "(...) and the HCR-20 v3 predicted violent offenses better than the HCR-20 v3 combined with the FAM (z = 2.504, p = .012, d = .220).*"

Comments to the Author: 4 Discussion
In general: A more in-depth and better organized discussion of the results, and comparisons with previous findings, is needed.

Authors' reply: *We have extensively revised the Discussion and now compare the results (recidivism rates and AUC values) with existing studies, consider the clinical implications of our findings, and make suggestions for future studies.*

Comments to the Author: First paragraph: Please compare the present rate of recidivism first to comparable studies in female forensic psychiatric patients, for instance the study of de Vogel et al. (23). You may analyze whether these figures differ significantly from your findings. If so, may specific characteristics of your patient group explain that (like high prevalence of substance abuse)? Then compare it to recidivism rates in male patients, etc.

Authors' reply: *We have added the following paragraph to the Discussion: "These recidivism rates are similar to those observed in the studies by de Vogel et al. [21] and Schaap et al. [43]. De Vogel et al. [21] examined a sample of 71 women who were discharged from forensic psychiatric hospitals. After a mean follow-up period of 11.8 years (SD = 4.9), 24 (33.8%) were officially reconvicted for one or more offenses, and in 13 (18.3%) cases, these offenses were violent. Schaap et al. [43] analyzed recidivism in 45 forensic inpatients and found that 16 (36%) were reconvicted of an offense (i.e., general*

recidivism) and 7 (16%) were reconvicted for a violent offense. We found not significant difference between these recidivism rates and those in the present sample (present study vs de Vogel et al.: $Chi^2(1) = 2.24$, $p = .135$, Cramer V = .0985; present study vs Schaap et al.: $Chi^2(1) = .39$, $p = .532$, Cramer V = .051)."

Comments to the Author: First paragraph, last sentence: Please corroborate this suggestion with literature.

Authors' reply: *The comparison with male patients was moved to Supplement 2, and literature was added here: "It is possible that the effect described in the literature (see 7,21) only applies to mentally healthy female offenders."*

Comments to the Author: Second paragraph: I miss comparisons with previous studies in female patients on various risk assessment instruments.

Authors' reply: *We have added a comparison (see pages 20-21): "When we compared the AUC values in our study with the prognostic validity of the risk assessment instruments HCR-20 v3, FAM, and PCL-R as reported by de Vogel et al. [21], we found similar, good metrics for predicting general recidivism (HCR-20 v3 = .667; HCR-20 v3 + FAM = .661 PCL-R = .601), but much better metrics in the present sample for predicting recidivism with a violent offense (HCR-20 v3 = .672; HCR-20 v3 + FAM = .651; PCL-R = .591). The better prognostic validity for recidivism with a violent offense in the present study may be because of the larger sample size. Generally, the AUC is not directly dependent on sample size, but it can indirectly be affected by it, especially when the sample is too small to contain a sufficient number of events needed for a reliable estimation of model performance. For example, at a recidivism rate of 34% (the general recidivism rate found by de Vogel et al.), a total sample size of 60 people is sufficient to test a prognostic instrument with an AUC of .714 against the null hypothesis (AUC = .5). However, at a recidivism rate of 18% (the violent recidivism found by de Vogel et al.), a total sample size of 80 people would be needed [44]."*

Comments to the Author: Table 5 can be omitted, as it does not add content.

Authors' reply: *Thank you for the suggestion. We have removed the table and restructured the Discussion accordingly.*

Comments to the Author: Line 53: decreased instead of increased?

Authors' reply*: Yes, you are right. Thank you very much for your attentive reading.*

Comments to the Author: Page 15, first paragraph: What is the use of the Youden indexes for the reader? When applied, how would the accuracy of predicting recidivism look like?

Authors' reply: *Upon the recommendation of another reviewer, the calculation of the Youden Index was removed from the manuscript.*

Comments to the Author: Page 16, line 12: You remark that LSI-R requires an interview with the patient. That is also true for the PCL-R.

Authors' reply: *We now mention the PCL-R in the limitations: "Second, for the same reason, we were not able to apply the PCL-R and LSI-R in interview form."*

Comments to the Author: What do the results add to the clinical practice?

Authors' reply: *We have added the following recommendations to the Discussion (see page 21): "Therefore, the following recommendations can be derived for the use of these instruments in forensic psychiatric samples of women: The low sensitivity of the prognostic instruments means that they should not be used (solely) to make decisions about the timing of discharge from a forensic psychiatric hospital because they do not reliably classify patients who will relapse after discharge, so public safety may be compromised in some cases; however, the instruments can assist clinicians in developing risk management plans that can be used to reduce individual risk within the framework of therapeutic interventions or social-pedagogical support measures."*

- LSI-R is best suited for predicting general recidivism in female forensic psychiatric patients
- HCR-20 v3 is best suited for predicting violent recidivism in female forensic psychiatric patients
- All instruments exhibit low sensitivity and are not suitable as the sole basis for discharge decisions
- Study’s strength: large sample size, long observation period

Evaluation of whether commonly used risk assessment tools are applicable to women in forensic psychiatric institutions

Abstract

Objective: By providing a structured assessment of specific risk factors, risk assessment tools allow statements to be made about the likelihood of future recidivism in people who have committed a crime. These tools were originally developed for and primarily tested in men and are mainly based on the usual criminological background of men. Despite significant progress in the last decade, there is still a lack of empirical research on female offenders, especially female forensic psychiatric inpatients. To improve prognosis in female offenders, we performed a retrospective study to compare the predictive quality of the following risk assessment tools: PCL-R, LSI-R, HCR-20 v3, FAM, and VRAG-R.

Method: Data were collected from the information available in the medical files of 525 female patients who had been discharged between 2001 and 2017. We examined the ability of the tools to predict general and violent recidivism by comparing the predictions with information from the Federal Central Criminal Register.

Results: Overall, the prediction instruments had moderate to good predictive performance, and the study confirmed their general applicability to female forensic psychiatric patients.

Conclusion: The LSI-R proved to be particularly valid for general recidivism, and both, LSI-R and HCR-20 v3, for violent recidivism.

Keywords: violent recidivism, recidivism, female offenders, mentally ill offenders, risk prediction, recidivism prognosis

1 Introduction

Mental illness is prevalent among criminal justice populations, and studies indicate that approximately 10% to 15% of inmates have psychotic disorders, 20% to 30% experience depression and bipolar disorders, and over 50% have a substance use disorder [1,2]. Factors contributing to the relationship between mental disorders and criminal behavior include emotion dysregulation as a core feature of many psychiatric disorders; cognitive distortions, such as irrational beliefs or misinterpretations of social cues; and difficulties in regulating behavior, including impulsivity and poor impulse control [3–5]. Mental disorders increase not only the risk of committing a crime but also the risk of recidivism. Research by Ogilvie et al. [6] found that a higher percentage of mentally ill individuals than individuals without psychiatric diagnoses returned to court for all categories of offenses, including violent (20.5% vs 8.5%, respectively), nonviolent (60.3% vs 40.3%, respectively), and other minor offenses (61.7% vs 44.1%, respectively). Co-occurring substance use disorders further exacerbate recidivism rates [7]. Furthermore, Okamura et al. [8] demonstrated that individuals who recidivate often lack medical services and support.

Mentally ill justice-involved individuals, both women and men, represent a distinct and vulnerable group within the criminal justice system and require specific and focused care [9]. They need access to customized psychosocial services that address their particular diagnoses, and they can benefit from psychotherapy and medication as part of their treatment. In addition, appropriate reintegration support after discharge is essential to enable successful rehabilitation and reduce the risk of relapse. Many countries aim to place mentally ill offenders in specialized facilities, forensic psychiatric ones, rather than in prison.

In forensic psychiatry, risk assessment tools are frequently used to guide decisions related to supervision and treatment and to assess the probability of reoffending [10]. These tools quantify the level of risk associated with a particular situation or individual by means of a structured assessment of specific, recidivism-related risk factors (such as age, previous

convictions, or social support) [11]. Recently, Ogonah et al. [12] conducted a meta-analysis to evaluate the performance of these tools in forensic mental health contexts. They reported pooled areas under the curve (AUCs) ranging from .64 to .74, predominantly in male samples (only two out of 50 studies assessed risk assessment tools in female-only samples). When they are applied to (mentally ill) justice-involved women, these tools may have the following significant limitations, which require careful consideration: (a) Justice-involved women are generally underrepresented in criminal justice data because they are only a small proportion (6.9%) of the overall offender population [13]; as a result of this underrepresentation, the data used to develop and validate risk assessment tools may be limited and less reliable [12,14]. (b) Justice-involved women often have distinct pathways into criminal behavior that differ from those of men, such as experiences with domestic violence [15–17]. (c) Justice-involved women tend to engage in different types of offenses than men and are more likely to be involved in non-violent crimes, such as drug-related offenses or property crimes, than in violent ones [18]; overall, research suggests that women generally have lower recidivism rates than men [14,19]. (d) Tools may underestimate or overlook the potential risks associated with mentally ill justice-involved women because not all of the distinctions observed between mentally healthy men and women necessarily extend to men and women with mental illness, e.g., although the risk of homicide and violent crime is lower in mentally healthy women than men, it is similar in mentally ill women and men [20,21].

To sum up, although risk assessment tools may have some usefulness in assessing future criminal behavior, they may not reliably assess risk in mentally ill justice-involved women. Underrepresentation in data, data bias, and the need for sex- and mental health-specific considerations all represent limitations in accurately predicting criminal behavior in mentally ill women. The present study aimed to examine various risk assessment tools and identify the most appropriate one in terms of the applicability to mentally ill justice-involved women.

2 Material and methods

2.1 Participants

In Germany, mentally ill offenders are admitted to a forensic psychiatric hospital on the basis of a court decision [22]. If a person has committed a serious crime because they have a serious mental disorder, such as schizophrenia, and if there is a high risk of recidivism, the court orders that the person be placed in the hospital in accordance with Section 63 of the German criminal code; for this group of patients, the length of hospitalization is not limited by law, but the longer the placement lasts, the more weight is given to the proportionality test, i.e., the risk of serious recidivism versus the patient's right to liberty. According to the German criminal code, a crime or recidivism is serious when the victims of the offence experience or are exposed to a considerable danger of severe emotional trauma or physical injury or the crime or recidivism causes serious economic damage. In contrast, offenders who committed a crime while under the influence of a substance use disorder are placed in a forensic psychiatric hospital in accordance with Section 64 of the German criminal code; in these individuals, hospitalization is usually for a maximum of two years, provided there is a high risk of recidivism and a favorable treatment prognosis. If, after discharge, people admitted according to Section 64 no longer meet the criteria for successful treatment, they may be returned to prison. Patients admitted under Section 63 or 64 are treated in specialized secure hospitals, where they are cared for by doctors, psychologists, and nurses rather than being supervised by security personnel. Currently, forensic psychiatric treatment focuses on addressing individual risk factors, such as specific symptoms and behaviors related to the offense, with the aim to minimize the risk of reoffending.

In the present study, data were collected from the records of 557 female forensic psychiatry patients at the Department of Forensic Psychiatry and Psychotherapy in Taufkirchen, Germany. Of these patients, 32 were excluded from further analysis because they had died (n = 13) or did not meet the inclusion criteria (n = 19). The inclusion criteria

required the presence of at least one written court judgment or one written psychiatric or psychological expert report. Thus, complete data sets were available for a total of 525 patients. The data were retrospectively analyzed by performing a risk assessment with the data in the patient records. All patients had been detained under either Section 63 (severe mental disorder, *n* = 110, 21%) or Section 64 (substance use disorder, *n* = 415, 79%) of the German criminal code and were discharged from the hospital between January 1, 2001, and December 31, 2017.

2.2 Risk assessments

We performed a literature review of 200 articles on criminal prognosis in women. *Prognosis* refers to the likely course and outcome of a condition, whereas *diagnosis* involves identifying the nature of an illness or problem through examination of the symptoms. The following five different assessment tools were deemed suitable for evaluation [23]: Psychopathy Checklist – Revised, PCL-R [24], Level of Service Inventory – Revised, LSI-R [25], Historical Clinical Risk Management-20, version 3, HCR-20 v3 [26], Female Additional Manual, FAM [27] as an extension of HCR-20 v3, and Violence Risk Appraisal Guide – Revised, VRAG-R [28]. In criminal prognosis, a distinction is made between actuarial risk assessment and the structured professional judgment approach: The former provides standardized and quantitative risk predictions based on statistical models (e.g., LSI-R, VRAG-R), whereas the latter offers a more flexible and nuanced approach that incorporates clinical expertise and case-specific information for assessing and managing risks (e.g., HCR-20 v3, FAM).

2.2.1 PCL-R

The PCL-R [24] is a third-party assessment of psychopathy personality traits (German translation by Mokros et al. [29]). It was not originally designed as a risk assessment tool, but it is commonly used as a prognostic tool because of its ability to predict recidivism. It assesses

20 items on a three-point rating scale (0 = definitely not present, 1 = not enough or inconsistent information to score the item, 2 = definitely present) that are assigned to two subcomponents: Factor 1, which includes features such as superficial charm, manipulation, and a grandiose self-image, and Factor 2, which encompasses antisocial behavior and an unstable lifestyle. In addition a total score is calculated, with higher values indicating higher expressed psychopathy traits. The scale is reported to have good psychometric properties (Cronbach's $\alpha = .87$, inter-rater reliability $r = .91$, and test-retest reliability $r = .94$, [30–32]). The predictive validity for general and violent relapses falls within the range of $r = .26$ to.28 [33], which can be considered to be moderate. In accordance with the descriptive schema for classifying PCL-R total scores [29, page 40], we chose the cut-off value of 24, which distinguishes between individuals with a moderate and those with a high to very high psychopathy score.

2.2.2 LSI-R

The LSI-R [25] was originally designed for use with probationers and parolees. However, it has also proven useful in other community corrections samples and in prisons, jails, halfway houses, and forensic mental health clinics and hospitals. The tool assesses the risk of reoffending by identifying criminogenic needs (German translation by Dahle et al. [34]). In total, 54 information areas are assessed, which are classified into ten different risk areas, referred to as "need scales." Items were rated according the manual. A higher total score indicates a higher risk of reoffending.

Interrater reliabilities are reported to be high ($r = .80$ to .94), and internal consistencies are moderate to high ($r = .41$ to.69) [34]. A meta-analysis found a predictive validity of $r = .35$ to .38 [25]. For short to medium prediction periods, the scale has good predictive validity for the risk of general recidivism ($r = .43$) [34]. In accordance with the LSI-R recidivism risk classification [34, page 72], we chose the cut-off value of 33, which distinguishes between

individuals with a moderate and those with an increased or high risk. Two issues related to the LSI-R must be mentioned here: First, the LSI-R is intended to be completed during an interview, which was not possible in the context of the present study, and second, some of the LSI-R items were difficult to apply to the conditions of a forensic psychiatric facility, so we had to adapt them for the present study.

2.2.3 HCR-20 v3

The HCR-20 v3 [26] uses a structured professional judgment approach to predict future violence and develop risk management strategies for forensic psychiatric patients (German translation by Bolzmacher [35]). Prognosis is assessed on the basis of 20 risk factors, which are subdivided into a historical scale (10 risk factors), clinical scale (5 risk factors), and risk management scale (5 risk factors). The median interrater reliability for the latest version of the HCR-20 (version 3) is reported to be .65 [36].

The use of a structured professional judgement approach means that the scale is not actually designed to quantify items, including summing the fulfilled risk factors. Instead, it relies on the professional's experience and subjective interpretation of the data. According to the HCR-20 v3 manual, the professional judgment consists of seven steps: gathering case information, assessing risk factors, assessing the relevance of risk factors, risk conceptualization, risk scenarios, risk management strategies, and final judgment. In the present study, the risk assessment was based on records, so we were not able to carry out steps 3 to 7. In accordance with Brookstein [37], we assessed the presence of risk factors (step 2) by using a 3-point scale (present = 2, partially present = 1, not present = 0) and summed the scores of the 20 risk factors, which yielded a total score ranging from 0 to 40.

2.2.4 FAM

FAM was developed as an extension of HCR-20 v3 to predict the relapse risk for violence in women and considers eight additional items [27]. For the present study, the English version of the FAM was used. The manual states that the eight additional items should be rated on a five-point scale. Like the HCR-20 v3, the FAM also follows the structured professional judgement approach, meaning that ratings are not intended to be given numerical values. However, to ensure better comparability of findings in the present study, we rated the items by using an approach similar to that used for the HCR-20 v3, i.e., on a three-point scale (0 = no, 1 = possible or partial, 2 = yes), and summed the scores of the 10 FAM risk factors, which yielded a total score ranging from 0 to 20. The authors reported good interrater reliabilities for all FAM items, with intraclass correlation coefficients (ICCs) ranging from .63 to .97 [27].

Because the combined evaluation of the HCR-20 v3 and FAM included 18 items from the HCR-20 v3 and 20 items from the FAM, the total scores ranged from 0 to 56.

2.2.5 VRAG-R

VRAG-R [28] is an actuarial assessment tool used to predict violence relapses (German version by Rettenberger [38]). It rates twelve items with different scoring systems, as described in the VRAG-R manual. The predictive validity for violent relapses is reported to be good (AUC, .76) [38]. In accordance with the risk categories of the VRAG-R [38, page 5], we chose the cut-off value of 11, which distinguishes between individuals with recidivism rates below 45% and equal to or greater than 45% after 5 years and below 69% and equal to or greater than 69% after 12 years.

2.3 Procedures

Before the study, five research staff members (clinical psychologists) were trained in the prognostic instruments. Then, the staff members assessed patients by referring to the patient

medical records and coded the final risk judgements. To confirm a uniform standard of assessment ratings across reviewers, all five reviewers independently rated 11 patients with the five assessment tools and interrater reliabilities were calculated (the ICCs of the assessment tools were as follows: PCL-R, .71; LSI-R, .74; HCR-20 v3, .61; FAM, .82; and VRAG-R, .89). To evaluate actual relapses after patients had been discharged, we obtained information from the German Federal Central Criminal Register in September 2020 and February 2021, in which all formal convictions are documented. Each formal conviction documented after release from the hospital or prison (in the case of treatment discontinuation) was counted as recidivism. In addition, violent offenses, which were defined as convictions for an offense involving crimes against persons (e.g., homicide, sex crimes, assault, threat, and robbery), were analyzed separately. The time at risk began at the time of release from the forensic psychiatric hospital (or prison) and ended when another crime was committed. If no further crime was committed, the time at risk ended on the date when the report was obtained from the German Federal Central Criminal Register. The mean time at risk was 6.0 years (standard deviation [SD], 4.9 years).

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2013. All procedures involving patients were approved by the ethics committee of the Bavarian Medical Association, Germany (approval no.: 2019-167). Informed consent was not necessary because of the retrospective nature of the study. This was approved by the ethics committee.

2.4 Statistical analysis

Means, SDs, and absolute and relative frequencies were calculated to describe the sample. To test for group differences between recidivists and non-recidivists, we used the Mann-Whitney *U* test.

Predictive validity was determined by dichotomizing the assessment scores. We used the cut-off values recommended in the manuals (see Methods section), as follows: PCL-R, > 24; LSI-R, > 33; and VRAG-R, > 11. We included the following commonly used primary outcome values: sensitivity, specificity, positive predictive value, negative predictive value, *r* (point-biserial correlation coefficient), and the AUC statistics [39]. Sensitivity indicates how reliably the assessment instrument correctly predicted relapse in patients who relapsed, and specificity indicates how reliably it correctly predicted the absence of relapse in patients who did not relapse. Positive predictive values were defined as the proportion of patients classified as high risk who went on to (violently) reoffend, and negative predictive values, as the proportion of patients classified as low risk who did not go on to (violently) reoffend. Point-biserial correlation analysis is a statistical procedure for measuring and evaluating the strength and direction (i.e., positive versus negative correlation) of the relationship between two variables. It allows one to determine whether changes in one variable are associated with changes in another and quantifies the degree of this connection (strong effect, ± .50 and above; medium effect, between ± .30 and ± .49; small effect, ± .29 and below [40]). The AUC statistics were determined with a receiver operation characteristics (ROC) curve, which is the function of the rate of true positives (i.e., sensitivity) and the rate of false positives (i.e., 1-specificity). The AUC expresses the accuracy of a prognostic tool in discriminating between relapsed and non-relapsed patients. An AUC of .5 indicates chance-level accuracy. According to commonly accepted standards, AUC values greater than .714 are generally considered to indicate good prognostic instrument performance [41]. To determine whether the differences between the AUCs of the assessment tools were statistically significant, the Mann-Whitney *U* test was used.

Finally, we compared the predictive performance of the five instruments in two subgroups: patients with schizophrenia (n = 81) and patients with a substance use disorder (n = 393). To test for group differences between recidivists and non-recidivists, we used the

Mann-Whitney *U* test. In addition, we calculated the AUC statistics for both subgroups and compared the AUCs with the non-parametric Mann-Whitney *U* test.

2.5 Transparency and openness

We report how we determined our sample size, all data exclusions, and all measures in the study, and we follow JARS [42]. All data and analysis codes are available from the corresponding author, [JS], upon reasonable request. Data were analyzed with IBM SPSS Statistics for Windows version 26 (Armonk, NY: IBM Corp.). This study's design and its analysis were not pre-registered.

3 Results

3.1 Sociodemographic characteristics

Table 1 shows the detailed sociodemographic information of the sample and Table 2 lists the forensic psychiatric characteristics of the sample. All patients were female. The most common diagnoses were substance use disorder (n = 393) and schizophrenia (n = 81). About 15% of the patients had a comorbid personality disorder. 40% of the patients (n = 208) relapsed, 11% (n = 60) with a violent offense.

Table 1

Sociodemographic information of the patients (N =525)

		M (*SD*)
Age (at hospital admission)		34.15 (10.14)
		Frequency (%)
Marital status		
	Single	302 (58%)
	Married / In a registered partnership	73 (14%)
	Widowed	15 (3%)

Separated / Divorced	135 (26%)
School and vocational training according to the International Standard Classification of Education	
No education	6 (1%)
Primary education	50 (10%)
Lower secondary education	233 (44%)
Upper secondary education	209 (40%)
Post-secondary non-tertiary education	10 (2%)
Tertiary education	16 (3%)
Occupation[a]	
Unemployed	410 (78%)
Employed	66 (13%)
Undergoing training	5 (1%)
Not capable of being employed	43 (8%)
Provision of parental care[b]	
No provision of parental care (despite having children)	128 (54%)
Sole parental caregiver	49 (21%)
Joint parental caregiver	61 (26%)

Note. [a]missing data = 1; [b]patients without children were not considered; *M*, mean; *SD*, standard deviation

Table 2

Forensic psychiatric characteristics of the patients (N = 525)

	M (*SD*)
Age at first crime (in years)[a]	23.94 (11.24)
Age at first inpatient treatment (in years)[b]	27.17 (10.73)
	Frequency (%)
Index offense	
Offense against public order	1 (.2%)
Sexual assault	1 (.2%)
Insult	1 (.2%)
Traffic offense	19 (4%)
Financial crime / Property damage	75 (14%)

Resistance against state authority	2 (.4%)
Coercion	9 (2%)
Robbery	29 (6%)
Drug-related crime	201 (38%)
Arson	29 (6%)
Assault	112 (21%)
Homicide	46 (9%)
Main clinical diagnosis and comorbid personality disorder[bc]	
F0: Organic disorder	4 (1%)
F10: Alcohol-related disorder	50 (10%)
F10 + F6: Alcohol-related disorder and personality disorder	21 (4%)
F11-18: Substance-related disorder to specific substance	118 (23%)
F11-18 + F6: Substance-related disorder to specific substance and personality disorder	10 (2%)
F19: Multiple drug use	157 (30%)
F19 + F6: Multiple drug use and personality disorder	37 (7%)
F2: Schizophrenic disorder	73 (14%)
F2 + F6: Schizophrenic disorder and personality disorder	8 (2%)
F3: Mood disorder	4 (.7%)
F3 + F6: Mood disorder and personality disorder	1 (.2%)
F4: Adjustment disorder / PTSD	1 (.2%)
F4 + F6: Adjustment disorder/PTSD and personality disorder	1 (.2%)
F6: Personality disorder	33 (6%)
F7: Mental retardation	1 (.2%)
F9: Conduct disorder	1 (.2%)
F9: Conduct disorder and personality disorder	2 (.4%)

Note. [a]missing data = 1; [b]missing data = 3; [c]diagnoses according to International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10), the diagnoses were made by physicians working in the forensic psychiatric hospital; PTSD, posttraumatic stress disorder; *M*, mean; *SD*, standard deviation

3.2 Predictive validity of assessment tools

Table 3 presents the metrics of the five predictive instruments regarding their prognostic validity for general recidivism. All instruments significantly distinguished between the group of patients with a recidivism offense and those without (see column test statistics). The confidence intervals of the AUC values for all five instruments encompassed the threshold of .714, indicating that none exceeded this value. Therefore, the prognostic validity can be interpreted as moderate to good. When comparing the AUC values of the instruments, the LSI-R proved to be the best instrument for predicting general recidivism in the present sample (LSI-R compared to PCL-R: $z = 2.956$, $p = .003$, $d = .260$; to HCR-20 v3: $z = 3.442$, $p = .001$, $d = .304$; to HCR-20 v3 and FAM: $z = 2.670$, $p = .008$, $d = .235$; and to VRAG-R: $z = 2.735$, $p = .006$, $d = .240$). Furthermore, the prediction of general recidivism with the combined HCR-20 v3 and FAM was not better than that with the HCR-20 v3 alone ($z = 1.780$, $p = .075$, $d = .156$). When considering the scales of the PCL-R, Factor 2 was a significantly better predictor of general recidivism than Factor 1 ($z = 4.298$, $p < .001$, $d = .382$). The HCR-20 v3 risk management subscale also performed significantly better than the HCR-20 v3 clinical scale ($z = 3.474$, $p = .001$, $d = .307$). All other pairwise comparisons of the AUC metrics not mentioned did not differ significantly from each other (AUC values at the item level can be found in Supplement 1).

Table 3

Means, standard deviations, test statistics of the Mann-Whitney U test comparing non-recidivists and recidivists, correlations, and area under the curve and confidence interval values of the assessment tools for general reoffending in the sample (N = 525)

	General reoffending			
	Non-recidivists	Recidivists	Test statistics	*r*, AUC, CI

		M (*SD*)	*M* (*SD*)		
PCL-R					
	Total score	12.54 (7.14)	17.42 (6.97)	*z* = 7.364*** *d* = .678	*r* = .320*** AUC = .690*** CI = .644, .735
	Factor 1	4.25 (3.52)	5.33 (3.49)	*z* = 3.588*** *d* = .317	*r* = .157*** AUC = .592*** CI = .543, .641
	Factor 2	7.09 (4.37)	10.47 (4.52)	*z* = 8.032*** *d* = .749	*r* = .351*** AUC = .707*** CI = .661, .752
LSI-R		21.97 (7.97)	29.37 (7.44)	*z* = 9.639*** *d* = .927	*r* = .423*** AUC = .748*** CI = .707, .790
HCR-20 v3					
	Total score	20.74 (7.09)	25.26 (6.21)	*z* = 7.102*** *d* = .651	*r* = .312*** AUC = .683*** CI = .637, .729
	Historical	12.82 (3.28)	14.38 (2.60)	*z* = 5.545*** *d* = .499	*r* = .242*** AUC = .642*** CI = .595, .690
	Clinical	2.84 (2.53)	3.98 (2.56)	*z* = 5.018*** *d* = .449	*r* = .219*** AUC = .628*** CI = .579, .677
	Risk managment	5.08 (2.77)	6.90 (2.47)	*z* = 7.411*** *d* = .684	*r* = .324*** AUC = .690*** CI = .644, .738
HCR-20 v3 and FAM		26.98 (9.19)	33.44 (8.23)	*z* = 7.737*** *d* = .717	*r* = .338*** AUC = .699*** CI = .654, .744
VRAG-R		-4.40 (16.79)	7.30 (15.45)	*z* = 7.516*** *d* = .694	*r* = .332*** AUC = .694*** CI = .649, .739

Note. PCL-R, Psychopathy Checklist – Revised; LSI-R, Level of Service Inventory – Revised; HCR-20 v3, Historical Clinical Risk Management-20, version 3; FAM, Female Additional Manual; VRAG-R, Violence Risk Appraisal Guide – Revised; *M*, mean; *SD*, standard deviation; *z*, standardized test statistic of the Mann-Whitney *U* test; *d*, Cohen's *d* effect size for unequal sample sizes; *r*, point-biserial correlation; AUC, area under curve; CI, 95% confidence interval; ***$p < .001$, **$p < .01$ (two-tailed tested)

Table 4 show the sensitivities, specificities, positive and negative predictive values of the assessment tools PCL-R, LSI-R, and VRAG-R for general reoffending. All three instruments had very low sensitivity, which was due to the fact that only a few patients were correctly classified as positive (PCL-R, 14%; LSI-R, 35%; VRAG-R, 45%). The specificity was good, with a large proportion of patients correctly classified as negative (PCL-R, 94%; LSI-R, 91%; VRAG-R, 81%).

Table 4

Sensitivities, specificities, positive and negative predictive values of three assessment tools for general reoffending in the sample (N = 525)

	General reoffending			
	Sensitivity[1]	Specificity[2]	PPV[3]	NPV[4]
PCL-R	.14	.94	.63	63
LSI-R	.35	.91	.72	.68
VRAG-R	.45	.81	.60	.69

Note. PCL-R, Psychopathy Checklist – Revised, cut-off value > 24; LSI-R, Level of Service Inventory – Revised, cut-off value > 33; VRAG-R, Violence Risk Appraisal Guide – Revised, cut-off value > 11; [1]Sensitivity = Cp/(Cp + Fn); [2]Specificity = Cn/(Cn + Fp); [3]Positive predictive value = Cp/(Cp + Fp); [4]Negative predictive value = Cn/(Cn + Fn) (Cp, number of correct positive outcomes; Cn, number of correct negative outcomes; Fp, number of false positive outcomes; Fn number of false negative outcomes; PPV, positive predictive value; NPV, negative predictive value); as structured professional judgement tools, HCR-20 v3 and FAM do not specify cut-off scores for classifying assessed individuals into different risk levels. Therefore, specificity, sensitivity, positive predictive value, and negative predictive valuecould not be calculated.

Table 5 presents the metrics for predicting recidivism with a violent offense. Here, too, all methods can distinguish between recidivist and non-recidivist patients (see column test statistics). Again, the confidence intervals of the AUC values of all five prognostic instruments included the threshold of .714, so all instruments can be considered to be moderate to good. The comparison of the AUC metrics of the instruments showed the following differences: The VRAG-R predicted violent offenses less well than the HCR-20 v3

(z = 2.397, p = .017, d = .210) and the LSI-R (z = 2.074, p = .038, d = .182), and the HCR-20 v3 predicted violent offenses better than the HCR-20 v3 combined with the FAM (z = 2.504, p = .012, d = .220). All other pairwise comparisons of the AUC metrics between the instruments or the scales were not significant.

Table 5

Means, standard deviations, and test statistics of the Mann-Whitney U test comparing non-recidivists and recidivists and r, area under the curve, and confidence interval values of the assessment tools for violent reoffending in the sample (N = 525)

		Violent reoffending			
		Non-recidivists *M* (*SD*)	Recidivists *M* (*SD*)	Test statistics	*r*, AUC, CI
PCL-R					
	Total score	13.79 (7.22)	19.77 (7.24)	z = 5.654*** d = .509	r = .255*** AUC = .724*** CI = .656, .791
	Factor 1	4.47 (3.50)	6.28 (3.55)	z = 3.734*** d = .330	r = .163*** AUC = .647*** CI = .578, .717
	Factor 2	7.99 (4.52)	11.80 (4.95)	z = 5.663*** d = .510	r = .247*** AUC = .724*** CI = .650, .798
LSI-R		24.07 (8.32)	31.37 (7.66)	z = 6.096*** d = .552	r = .272*** AUC = .741*** CI = .675, .808
HCR-20 v3					
	Total score	21.80 (6.96)	28.17 (5.52)	z = 6.654*** d = .606	r = .285*** AUC = .764*** CI = .703, .824
	Historical	13.16 (3.08)	15.57 (2.51)	z = 5.870*** d = .530	r = .256*** AUC = .732*** CI = .666, .797
	Clinical	3.09 (2.57)	4.88 (2.34)	z = 5.114***	r = .223***

			d = .458	AUC = .701*** CI = .636, .766
	Risk managment	5.56 (2.77)	7.72 (2.18)	*z* = 5.847*** *d* = .528	*r* = .255*** AUC = .730*** CI = .667, .794
HCR-20 v3 and FAM		29.35 (9.72)	36.10 (7.88)	*z* = 5.766*** *d* = .520	*r* = .252*** AUC = .728*** CI = .665, .792
VRAG-R		-1.02 (17.00)	9.94 (15.99)	*z* = 4.571*** *d* = .407	*r* = .203*** AUC = .681*** CI = .611, .752

Note. PCL-R, Psychopathy Checklist – Revised; LSI-R, Level of Service Inventory – Revised; HCR-20 v3, Historical Clinical Risk Management-20, version 3; FAM, Female Additional Manual; VRAG-R, Violence Risk Appraisal Guide – Revised; *M*, mean; *SD*, standard deviation; *z*, standardized test statistic of the Mann-Whitney *U* test; *d*, Cohen's *d* effect size for unequal sample sizes; *r*, point-biserial correlation; AUC, area under curve; CI, 95% confidence interval, ****p* < .001, ***p* < .01 (two-tailed tested)

Table 6 displays the sensitivities, specificities, and positive and negative predictive values of the assessment tools PCL-R, LSI-R, and VRAG-R for predicting violent reoffending. The sensitivity of the three instruments for predicting violent recidivism was very low (correctly classified as positive: PCL-R, 14%; LSI-R, 35%; VRAG-R, 45%). Overall, 143 of 525 (27%) patients recidivated with a violent offense without recidivism being predicted.

Table 6

Sensitivities, specificities, positive and negative predictive values of three assessment tools for violent reoffending in the sample (N = 525)

	Violent reoffending			
	Sensitivity[1]	Specificity[2]	PPV[3]	NPV[4]
PCL-R	.22	.92	.27	.90
LSI-R	.48	.85	.29	.93
VRAG-R	.50	.73	.19	.92

Note. PCL-R, Psychopathy Checklist – Revised, cut-off value > 24; LSI-R, Level of Service Inventory – Revised, cut-off value > 33; VRAG-R, Violence Risk Appraisal Guide – Revised, cut-off value > 11; [1]Sensitivity = Cp/(Cp + Fn); [2]Specificity = Cn/(Cn + Fp); [3]Positive predictive value = Cp/(Cp + Fp);

[4]Negative predictive value = Cn/(Cn + Fn) (Cp, number of correct positive outcomes; Cn, number of correct negative outcomes; Fp, number of false positive outcomes; Fn number of false negative outcomes; PPV, positive predictive value; NPV, negative predictive value); as structured professional judgement tools, HCR-20 v3 and FAM do not specify cut-off scores for classifying assessed individuals into different risk levels. Therefore, specificity, sensitivity, positive predictive value, and negative predictive value could not be calculated.

In a further analysis, we compared the predictive performance of the five instruments in two subgroups: patients with schizophrenia (*n*=81) and patients with a substance use disorder (*n*=393). In patients with schizophrenia, relapse occurred in 19% (*n* = 15) and was characterized by a violent offense in 11% (*n* = 9); in patients with substance use disorder, it occurred in 45% (*n* = 176) and was characterized by a violent offense in 12% (*n* = 46). Compared with patients with schizophrenia, patients with substance use disorder relapsed significantly more often with a general offense ($Chi^2(1) = 19.257$; $p < .010$; Cramer $V = .202$). Regarding violent recidivism, no differences were found between the two diagnostic groups ($Chi^2(1) = .023$; $p = .879$; Cramer $V = .007$). The results revealed no significant differences between the two groups in terms of the predictive validity of the five assessment tools; the AUC differences ranged from -.007 to .106 and did not differ from zero.

4 Discussion

The present study aimed to examine prognosis of recidivism in women treated in forensic psychiatric facilities by evaluating the predictive quality of common prognostic instruments (PCL-R, LSI-R, HCR-20 v3, HCR-20 v3 + FAM, and VRAG-R). After a mean observation period of 6.0 years (SD = 4.9 years), general recidivism had occurred in 208 (40%) of the 525 women examined, and violent recidivism in 60 (11%). These recidivism rates are similar to those observed in the studies by de Vogel et al. [21] and Schaap et al. [43]. De Vogel et al. [21] examined a sample of 71 women who were discharged from forensic psychiatric hospitals. After a mean follow-up period of 11.8 years (*SD* = 4.9), 24 (33.8%) were officially reconvicted for one or more offenses, and in 13 (18.3%) cases, these offenses were violent.

Schaap et al. [43] analyzed recidivism in 45 forensic inpatients and found that 16 (36%) were reconvicted of an offense (i.e., general recidivism) and 7 (16%) were reconvicted for a violent offense. We found not significant difference between these recidivism rates and those in the present sample (present study vs de Vogel et al.: $Chi^2(1) = 2.240$, $p = .135$, Cramer $V = .098$; present study vs Schaap et al.: $Chi^2(1) = .391$, $p = .532$, Cramer $V = .051$). For a comparison of the recidivism rates in the present female sample with a comparable male sample of forensic psychiatric inpatients, see Supplement 2.

The current study shows that the evaluated risk assessment tools are suitable for use in female forensic psychiatric patients because the tools were able to reliably differentiate between patients with and without general and violent recidivism. With regard to the quality of the predictions (AUC), the predictions were significantly better than chance, and the instruments consistently showed moderate to good performance. For general recidivism, the LSI-R had the best predictive quality, and for violent recidivism, the HCR-20 v3 and LSI-R both performed well. The present study further showed that supplementing the HCR-20 v3 with the FAM does not improve the prognosis, neither for the prediction of general recidivism nor for the prediction of violent recidivism. Thus, our data indicate that supplementary use of the FAM is not helpful in predicting recidivism.

When we compared the AUC values in our study with the prognostic validity of the risk assessment instruments HCR-20 v3, FAM, and PCL-R as reported by de Vogel et al. [21], we found similar, good metrics for predicting general recidivism (HCR-20 v3 = .667; HCR-20 v3 + FAM = .661 PCL-R = .601), but much better metrics in the present sample for predicting recidivism with a violent offense (HCR-20 v3 = .672; HCR-20 v3 + FAM = .651; PCL-R = .591). The better prognostic validity for recidivism with a violent offense in the present study may be because of the larger sample size. Generally, the AUC is not directly dependent on sample size, but it can indirectly be affected by it, especially when the sample is too small to contain a sufficient number of events needed for a reliable estimation of model

performance. For example, at a recidivism rate of 34% (the general recidivism rate found by de Vogel et al.), a total sample size of 60 people is sufficient to test a prognostic instrument with an AUC of .714 against the null hypothesis (AUC = .5). However, at a recidivism rate of 18% (the violent recidivism found by de Vogel et al.), a total sample size of 80 people would be needed [44].

In a further analysis, we examined whether predictive accuracy was influenced by patient diagnosis and found that it was not, i.e., all five prognostic instruments achieved comparably good results in both subgroups. Further differentiation of the instruments depending on individual diagnoses does not appear to be necessary when using the instruments in mentally ill women.

In the present study all prognostic instruments had rather low sensitivity for predicting general and violent offending. To improve sensitivity, the cut-off values could be decreased, although that usually results in a slight decrease in specificity. Tools with high sensitivity will be most effective at safeguarding the public and may also gain significant political support. However, tools with high specificity will best protect the rights and interests of psychiatric patients. Achieving a balance between false positives and false negatives is an ethical matter and depends on the social and political context in which the tool is being used. Therefore, the following recommendations can be derived for the use of these instruments in forensic psychiatric samples of women: The low sensitivity of the prognostic instruments means that they should not be used (solely) to make decisions about the timing of discharge from a forensic psychiatric hospital because they do not reliably classify patients who will relapse after discharge, so public safety may be compromised in some cases; however, the instruments can assist clinicians in developing risk management plans that can be used to reduce individual risk within the framework of therapeutic interventions or social-pedagogical support measures.

Future studies could examine various measures to enhance the sensitivity of instruments used in forensic psychiatric samples of women, such as altered cut-off scores, different weighting of individual risk factors, and the combined application of different instruments.

The present study has some limitations that should be considered when interpreting the results and drawing conclusions. First, this was a retrospective study in which the items of the prognostic instruments were rated with data collected from information in the files of patients who had already been discharged. As a result, missing data could not be ascertained retrospectively and the accuracy of the information could not be verified. Second, for the same reason, we were not able to apply the PCL-R and LSI-R in interview form. Third, the HCR-20 v3 is not designed for quantifying items, which limits the transferability of our results; however, the AUC value for predicting a violent offense based on the HCR-20 v3 total score was very good compared with that of the other instruments, suggesting that summing the risk factors yields excellent results, obviating the need to implement steps 3 to 7. Therefore, in routine clinical care, if professionals can only assess a patient based on records, they can achieve a good prognosis by summing the fulfilled factor values. Fourth, recidivism was assessed from entries in the Federal Central Register, so incidents from the dark figure of crime were not captured. Additionally, in accordance with Section 46 of the Federal Central Register Act (BZRG), entries in the Federal Central Register are deleted after the expiration of a specified time limit, which depends on the severity of the offense. Minor offenses (e.g., fines up to 90 daily rates) are deleted after 3 years, offenses resulting in a sentence of no more than one year of imprisonment are deleted after 5 years, and more serious offenses (e.g., sentences exceeding one year of imprisonment) are deleted after 10 years. And last, all instruments were scored by one scorer, so it is possible that the outcome of one instrument may have influenced the outcome of another.

5 Conclusions

A significant strength of the study lies in the large sample size (N = 525) and the extended observation period (mean time at risk, 6 years). Thus, the study yielded useful results on the prognostic validity and generalizability of the studied instruments. The AUC metrics indicate that all assessment instruments can be used to predict general and violent recidivism in women in forensic psychiatric care. In particular, the LSI-R appears to perform best in predicting general recidivism and both the HCR-20 v3 and LSI-R appear to perform equally well for the specific prediction of violent recidivism. All instruments exhibit low sensitivity and are not suitable as the sole basis for discharge decisions because they do not correctly classify a high proportion of patients who reoffend with a violent offense. Nevertheless, by highlighting individual risk areas, they can provide valuable information for planning therapy goals or support measures.

Funding

This work was supported by the Bavarian State Ministry for Family, Labor and Social Affairs, Germany [grant number ZBFS-X/1-10.700-5/3/9].

Acknowledgments

The codebook was designed in collaboration with the Office of Corrections and Rehabilitation, Zurich, Switzerland. The authors thank Jacquie Klesing, Board-certified Editor in the Life Sciences (ELS), for editing assistance with the manuscript.

Data statement

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

References

[1] Fazel S, Danesh J. Serious mental disorder in 23 000 prisoners: a systematic review of 62 surveys. The Lancet 2002;359:545–50. https://doi.org/10.1016/S0140-6736(02)07740-1.

[2] Fazel S, Hayes AJ, Bartellas K, Clerici M, Trestman R. Mental health of prisoners: prevalence, adverse outcomes, and interventions. Lancet Psychiatry 2016;3:871–81. https://doi.org/10.1016/S2215-0366(16)30142-0.

[3] Garofalo C, Velotti P. Negative emotionality and aggression in violent offenders: The moderating role of emotion dysregulation. J Crim Justice 2017;51:9–16. https://doi.org/10.1016/j.jcrimjus.2017.05.015.

[4] Newhill CE, Mulvey EP. Emotional dysregulation: The key to a treatment approach for violent mentally ill individuals. Clin Soc Work J 2002;30:157–71. https://doi.org/10.1023/A:1015293428307.

[5] Ullrich S, Keers R, Coid JW. Delusions, anger, and serious violence: new findings from the MacArthur Violence Risk Assessment Study. Schizophr Bull 2014;40:1174–81. https://doi.org/10.1093/schbul/sbt126.

[6] Ogilvie JM, Tzoumakis S, Thompson C, Allard T, Dennison S, Kisely S, et al. Psychiatric illness and the risk of reoffending: recurrent event analysis for an Australian birth cohort. BMC Psychiatry 2023;23:355. https://doi.org/10.1186/s12888-023-04839-0.

[7] Zgoba KM, Reeves R, Tamburello A, Debilio L. Criminal recidivism in inmates with mental illness and substance use disorders. J Am Acad Psychiatry Law 2020;48:209–15. https://doi.org/10.29158/JAAPL.003913-20.

[8] Okamura M, Okada T, Okumura Y. Recidivism among prisoners with severe mental disorders. Heliyon 2023;9. https://doi.org/10.1016/j.heliyon.2023.e17007.

[9] KiDeuk K, Becker-Cohen M, Serakos M. The Processing and Treatment of Mentally Ill Persons in the Criminal Justice System: A Scan of Practice and Background Analysis. Washington, DC: Urban Institute; 2015.

[10] Douglas T, Pugh J, Singh I, Savulescu J, Fazel S. Risk assessment tools in criminal justice and forensic psychiatry: The need for better data. Eur Psychiatry 2017;42:134–7. https://doi.org/10.1016/j.eurpsy.2016.12.009.

[11] Brown J, Singh JP. Forensic risk assessment: A beginner's guide. Arch Forensic Psychol 2014;1:49–59.

[12] Ogonah MG, Seyedsalehi A, Whiting D, Fazel S. Violence risk assessment instruments in forensic psychiatric populations: a systematic review and meta-analysis. Lancet Psychiatry 2023;10:780–9. https://doi.org/10.1016/S2215-0366(23)00256-0.

[13] Walmsley R. World female imprisonment list. Women and girls in penal institutions, including pre-trial detainees / remand prisoners. London: International centre for prison studies; 2017.

[14] Geraghty KA, Woodhams J. The predictive validity of risk assessment tools for female offenders: A systematic review. Aggress Violent Behav 2015;21:25–38. https://doi.org/10.1016/j.avb.2015.01.002.

[15] Katz RS. Explaining girl's and women's crime and desistance in the context of their victimization experiences – A developmental test of revised strain theory and the life course perspective. Violence Women 2000;6:633–60. https://doi.org/10.1177/1077801200006006005.
[16] van Voorhis P, Wright EM, Salisbury E, Bauman A. Women's risk factors and their contributions to existing risk/needs assessment. Crim Justice Behav 2010;37:261–88. https://doi.org/10.1177/0093854809357442.
[17] Johansson P, Kempf-Leonard K. A gender-specific pathway to serious, violent, and chronic offending? Exploring Howell's risk factors for serious delinquency. Crime Delinquency 2009;55:216–40. https://doi.org/10.1177/0011128708330652.
[18] Vogel V, Nicholls TL. Gender matters: An introduction to the special issues on women and girls. Int J Forensic Ment Health 2016;15:1–25. https://doi.org/10.1080/14999013.2016.1141439.
[19] Blanchett K, Brown SL. The assessment and treatment of women offenders: An integrative perspective. Chichester: John Wiley & Sons Ltd.; 2006.
[20] Nicholls TL, Brink J, Greaves C, Lussier P, Verdun-Jones S. Forensic psychiatric inpatients and aggression: an exploration of incidence, prevalence, severity, and interventions by gender. Int J Law Psychiatry 2009;32:23–30. https://doi.org/10.1016/j.ijlp.2008.11.007.
[21] Vogel V, Bruggeman M, Lancel M. Gender-sensitive violence risk assessment: Predictive validity of six tools in female forensic psychiatric patients. Crim Justice Behav 2019;46:528–49. https://doi.org/10.1177/0093854818824135.
[22] Müller JL, Saimeh N, Briken P, Eucker S, Hoffmann K, Koller M, et al. Standards für die Behandlung im Maßregelvollzug nach §§ 63 und 64 StGB. Forensische Psychiatr Psychol Kriminol 2018;12:93–125. https://doi.org/10.1007/s11757-017-0445-0.
[23] Mayer J, Wolf V, Steiner I, Franke I, Klein V, Streb J, et al. Rückfallprognose bei Straftäterinnen. Forensische Psychiatr Psychol Kriminol 2023;17:189–98. https://doi.org/10.1007/s11757-023-00770-y.
[24] Hare RD. Hare Psychopathy Checklist-Revised (PCL-R). Toronto: Multi-Health Systems; 2003.
[25] Andrews D, Bonta J. The Level of Service Inventory-Revised (LSI-R). Toronto: Multi Health Systems; 1995.
[26] Douglas KS, Hart SD, Webster CD, Belfrage H. HCR-20 V3: Assessing risk for violence – User guide. Simon Fraser University, Canada: Mental Health, Law, and Policy Institute; 2013.
[27] Vogel V, Vries Robbé M, van Kalmthout W, Place C. Female Additional Manual (FAM): Additional guidelines to the HCR-20 V3 for assessing risk for violence in women. Utrecht: Van der Hoeven Kliniek; 2014.
[28] Rice ME, Harris GT, Lang C. Validation of and revision to the VRAG and SORAG: the Violence Risk Appraisal Guide-Revised (VRAG-R). Psychol Assess 2013;25:951–65. https://doi.org/10.1037/a0032878.
[29] Mokros A, Hollerbach P, Nitschke J, Habermeyer E. PCL-R. Hare Psychopathy Checklist – Revised. Deutsche Version der Hare Psychopathy Checklist – Revised (PCL-R) von R. D. Hare. Göttingen: Hogrefe; 2017.
[30] Cooke DJ, Michie C. An item response theory evaluation of Hare's Psychopathy Checklist. Psychol Assess 1997;9:2–13. https://doi.org/10.1037/1040-3590.9.1.3.
[31] Hare RD, Harpur TJ, Hakstian AR, Forth AE, Hart SD, Newman JP. The revised psychopathy checklist: reliability and factor structure. Psychol Assess J Consult Clin Psychol 1990;2:338. https://doi.org/10.1037/1040-3590.2.3.338.

[32] Harris GT, Rice ME, Quinsey VL. Psychopathy as a taxon: evidence that psychopaths are a discrete class. J Consult Clin Psychol 1994;62:387. https://doi.org/10.1037/0022-006X.62.2.387.
[33] Dahle K-P, Schneider V, Ziethen F. Standardisierte Instrumente zur Kriminalprognose. Forensische Psychiatr Psychol Kriminol 2007;1:15–26. https://doi.org/10.1007/s11757-006-0004-6.
[34] Dahle K-P, Harwardt F, Schneider-Njepel V. Inventar zur Einschätzung des Rückfallrisikos und des Betreuungs- und Behandlungsbedarfs von Straftätern: LSI-R; deutsche Version des Level of Service Inventory-Revised nach Don Andrews und James Bonta. Göttingen: Hogrefe; 2012.
[35] Bolzmacher M, Born P, Eucker S, Franqué F, Holzinger B, Kötter S, et al. Die Vorhersage von Gewalttaten mit dem HCR-20 v3. Gießen: Institut für forensische Psychiatrie Haina e. V.; 2014.
[36] Franqué F. HCR-20 - Historical-Clinical-Risk Management-20 Violence Risk Assessment Scheme. In: Rettenberger M, Franqué F, editors. Handb. Kriminalprognostischer Verfahr., Göttingen: Hogrefe; 2013, p. 141–58.
[37] Brookstein DM, Daffern M, Ogloff JRP, Campbell RE, Chu CM. Predictive validity of the HCR-20 ^{V3} in a sample of Australian forensic psychiatric patients. Psychiatry Psychol Law 2021;28:325–42. https://doi.org/10.1080/13218719.2020.1775152.
[38] Rettenberger M, Gregório Hertz P, Eher R. Die deutsche Version des Violence Risk Appraisal Guide-Revised (VRAG-R). Wiesbaden: Kriminologische Zentralstelle; 2017.
[39] Rice ME, Harris GT. Violent recidivism: Assessing predictive validity. J Consult Clin Psychol 1995;63:737–48. https://doi.org/10.1037//0022-006x.63.5.737.
[40] Cohen J. Statistical power analysis for the behavioral sciences. routledge; 2013.
[41] Rice ME, Harris GT. Comparing Effect Sizes in Follow-Up Studies: ROC Area, Cohen's d, and r. Law Hum Behav 2005;29:615–20. https://doi.org/10.1007/s10979-005-6832-7.
[42] Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. Am Psychol 2018;73:3. https://psycnet.apa.org/doi/10.1037/amp0000389.
[43] Schaap G, Lammers S, Vogel V. Risk assessment in female forensic psychiatric patients: A quasi-prospective study into the validity of the HCR-20 and PCL-R. J Forensic Psychiatry Psychol 2009;20:354–65. https://doi.org/10.1080/14789940802542873.
[44] Lu G. Sample size formulas for estimating areas under the receiver operating characteristic curves with precision and assurance. Master's Thesis. The University of Western Ontario (Canada), 2021.

Evaluation of whether commonly used risk assessment tools are applicable to women in forensic psychiatric institutions

Abstract

Objective: By providing a structured assessment of specific risk factors, risk assessment tools allow statements to be made about the likelihood of future recidivism in people who have committed a crime. These tools were originally developed for and primarily tested in men and are mainly based on the usual criminological background of men. Despite significant progress in the last decade, there is still a lack of empirical research on female offenders, especially female forensic psychiatric inpatients. To improve prognosis in female offenders, we performed a retrospective study to compare the predictive quality of the following risk assessment tools: PCL-R, LSI-R, HCR-20 v3, FAM, and VRAG-R.

Method: Data were collected from the information available in the medical files of 525 female patients who had been discharged between 2001 and 2017. We examined the ability of the tools to predict general and violent recidivism by comparing the predictions with information from the Federal Central Criminal Register.

Results: Overall, the prediction instruments had moderate to good predictive performance, and the study confirmed their general applicability to female forensic psychiatric patients.

Conclusion: The LSI-R proved to be particularly valid for general recidivism, and both, LSI-R and HCR-20 v3, for violent recidivism.

Keywords: violent recidivism, recidivism, female offenders, mentally ill offenders, risk prediction, recidivism prognosis

1 Introduction

Mental illness is prevalent among criminal justice populations, and studies indicate that approximately 10% to 15% of inmates have psychotic disorders, 20% to 30% experience depression and bipolar disorders, and over 50% have a substance use disorder [1,2]. Factors contributing to the relationship between mental disorders and criminal behavior include emotion dysregulation as a core feature of many psychiatric disorders; cognitive distortions, such as irrational beliefs or misinterpretations of social cues; and difficulties in regulating behavior, including impulsivity and poor impulse control [3–5]. Mental disorders increase not only the risk of committing a crime but also the risk of recidivism. Research by Ogilvie et al. [6] found that a higher percentage of mentally ill individuals than individuals without psychiatric diagnoses returned to court for all categories of offenses, including violent (20.5% vs 8.5%, respectively), nonviolent (60.3% vs 40.3%, respectively), and other minor offenses (61.7% vs 44.1%, respectively). Co-occurring substance use disorders further exacerbate recidivism rates [7]. Furthermore, Okamura et al. [8] demonstrated that individuals who recidivate often lack medical services and support.

Mentally ill justice-involved individuals, both women and men, represent a distinct and vulnerable group within the criminal justice system and require specific and focused care [9]. They need access to customized psychosocial services that address their particular diagnoses, and they can benefit from psychotherapy and medication as part of their treatment. In addition, appropriate reintegration support after discharge is essential to enable successful rehabilitation and reduce the risk of relapse. Many countries aim to place mentally ill offenders in specialized facilities, forensic psychiatric ones, rather than in prison.

In forensic psychiatry, risk assessment tools are frequently used to guide decisions related to supervision and treatment and to assess the probability of reoffending [10]. These tools quantify the level of risk associated with a particular situation or individual by means of a structured assessment of specific, recidivism-related risk factors (such as age, previous

convictions, or social support) [11]. Recently, Ogonah et al. [12] conducted a meta-analysis to evaluate the performance of these tools in forensic mental health contexts. They reported pooled areas under the curve (AUCs) ranging from .64 to .74, predominantly in male samples (only two out of 50 studies assessed risk assessment tools in female-only samples). When they are applied to (mentally ill) justice-involved women, these tools may have the following significant limitations, which require careful consideration: (a) Justice-involved women are generally underrepresented in criminal justice data because they are only a small proportion (6.9%) of the overall offender population [13]; as a result of this underrepresentation, the data used to develop and validate risk assessment tools may be limited and less reliable [12,14]. (b) Justice-involved women often have distinct pathways into criminal behavior that differ from those of men, such as experiences with domestic violence [15–17]. (c) Justice-involved women tend to engage in different types of offenses than men and are more likely to be involved in non-violent crimes, such as drug-related offenses or property crimes, than in violent ones [18]; overall, research suggests that women generally have lower recidivism rates than men [14,19]. (d) Tools may underestimate or overlook the potential risks associated with mentally ill justice-involved women because not all of the distinctions observed between mentally healthy men and women necessarily extend to men and women with mental illness, e.g., although the risk of homicide and violent crime is lower in mentally healthy women than men, it is similar in mentally ill women and men [20,21].

To sum up, although risk assessment tools may have some usefulness in assessing future criminal behavior, they may not reliably assess risk in mentally ill justice-involved women. Underrepresentation in data, data bias, and the need for sex- and mental health-specific considerations all represent limitations in accurately predicting criminal behavior in mentally ill women. The present study aimed to examine various risk assessment tools and identify the most appropriate one in terms of the applicability to mentally ill justice-involved women.

2 Material and methods

2.1 Participants

In Germany, mentally ill offenders are admitted to a forensic psychiatric hospital on the basis of a court decision [22]. If a person has committed a serious crime because they have a serious mental disorder, such as schizophrenia, and if there is a high risk of recidivism, the court orders that the person be placed in the hospital in accordance with Section 63 of the German criminal code; for this group of patients, the length of hospitalization is not limited by law, but the longer the placement lasts, the more weight is given to the proportionality test, i.e., the risk of serious recidivism versus the patient's right to liberty. According to the German criminal code, a crime or recidivism is serious when the victims of the offence experience or are exposed to a considerable danger of severe emotional trauma or physical injury or the crime or recidivism causes serious economic damage. In contrast, offenders who committed a crime while under the influence of a substance use disorder are placed in a forensic psychiatric hospital in accordance with Section 64 of the German criminal code; in these individuals, hospitalization is usually for a maximum of two years, provided there is a high risk of recidivism and a favorable treatment prognosis. If, after discharge, people admitted according to Section 64 no longer meet the criteria for successful treatment, they may be returned to prison. Patients admitted under Section 63 or 64 are treated in specialized secure hospitals, where they are cared for by doctors, psychologists, and nurses rather than being supervised by security personnel. Currently, forensic psychiatric treatment focuses on addressing individual risk factors, such as specific symptoms and behaviors related to the offense, with the aim to minimize the risk of reoffending.

In the present study, data were collected from the records of 557 female forensic psychiatry patients at the Department of Forensic Psychiatry and Psychotherapy in Taufkirchen, Germany. Of these patients, 32 were excluded from further analysis because they had died (n = 13) or did not meet the inclusion criteria (n = 19). The inclusion criteria

required the presence of at least one written court judgment or one written psychiatric or psychological expert report. Thus, complete data sets were available for a total of 525 patients. The data were retrospectively analyzed by performing a risk assessment with the data in the patient records. All patients had been detained under either Section 63 (severe mental disorder, *n* = 110, 21%) or Section 64 (substance use disorder, *n* = 415, 79%) of the German criminal code and were discharged from the hospital between January 1, 2001, and December 31, 2017.

2.2 Risk assessments

We performed a literature review of 200 articles on criminal prognosis in women. *Prognosis* refers to the likely course and outcome of a condition, whereas *diagnosis* involves identifying the nature of an illness or problem through examination of the symptoms. The following five different assessment tools were deemed suitable for evaluation [23]: Psychopathy Checklist – Revised, PCL-R [24], Level of Service Inventory – Revised, LSI-R [25], Historical Clinical Risk Management-20, version 3, HCR-20 v3 [26], Female Additional Manual, FAM [27] as an extension of HCR-20 v3, and Violence Risk Appraisal Guide – Revised, VRAG-R [28]. In criminal prognosis, a distinction is made between actuarial risk assessment and the structured professional judgment approach: The former provides standardized and quantitative risk predictions based on statistical models (e.g., LSI-R, VRAG-R), whereas the latter offers a more flexible and nuanced approach that incorporates clinical expertise and case-specific information for assessing and managing risks (e.g., HCR-20 v3, FAM).

2.2.1 PCL-R

The PCL-R [24] is a third-party assessment of psychopathy personality traits (German translation by Mokros et al. [29]). It was not originally designed as a risk assessment tool, but it is commonly used as a prognostic tool because of its ability to predict recidivism. It assesses

20 items on a three-point rating scale (0 = definitely not present, 1 = not enough or inconsistent information to score the item, 2 = definitely present) that are assigned to two subcomponents: Factor 1, which includes features such as superficial charm, manipulation, and a grandiose self-image, and Factor 2, which encompasses antisocial behavior and an unstable lifestyle. In addition a total score is calculated, with higher values indicating higher expressed psychopathy traits. The scale is reported to have good psychometric properties (Cronbach's α = .87, inter-rater reliability *r* = .91, and test-retest reliability *r* = .94, [30–32]). The predictive validity for general and violent relapses falls within the range of *r* = .26 to.28 [33], which can be considered to be moderate. In accordance with the descriptive schema for classifying PCL-R total scores [29, page 40], we chose the cut-off value of 24, which distinguishes between individuals with a moderate and those with a high to very high psychopathy score.

2.2.2 LSI-R

The LSI-R [25] was originally designed for use with probationers and parolees. However, it has also proven useful in other community corrections samples and in prisons, jails, halfway houses, and forensic mental health clinics and hospitals. The tool assesses the risk of reoffending by identifying criminogenic needs (German translation by Dahle et al. [34]). In total, 54 information areas are assessed, which are classified into ten different risk areas, referred to as "need scales." Items were rated according the manual. A higher total score indicates a higher risk of reoffending.

Interrater reliabilities are reported to be high (*r* = .80 to .94), and internal consistencies are moderate to high (*r* = .41 to.69) [34]. A meta-analysis found a predictive validity of *r* = .35 to .38 [25]. For short to medium prediction periods, the scale has good predictive validity for the risk of general recidivism (*r* = .43) [34]. In accordance with the LSI-R recidivism risk classification [34, page 72], we chose the cut-off value of 33, which distinguishes between

individuals with a moderate and those with an increased or high risk. Two issues related to the LSI-R must be mentioned here: First, the LSI-R is intended to be completed during an interview, which was not possible in the context of the present study, and second, some of the LSI-R items were difficult to apply to the conditions of a forensic psychiatric facility, so we had to adapt them for the present study.

2.2.3 HCR-20 v3

The HCR-20 v3 [26] uses a structured professional judgment approach to predict future violence and develop risk management strategies for forensic psychiatric patients (German translation by Bolzmacher [35]). Prognosis is assessed on the basis of 20 risk factors, which are subdivided into a historical scale (10 risk factors), clinical scale (5 risk factors), and risk management scale (5 risk factors). The median interrater reliability for the latest version of the HCR-20 (version 3) is reported to be .65 [36].

The use of a structured professional judgement approach means that the scale is not actually designed to quantify items, including summing the fulfilled risk factors. Instead, it relies on the professional's experience and subjective interpretation of the data. According to the HCR-20 v3 manual, the professional judgment consists of seven steps: gathering case information, assessing risk factors, assessing the relevance of risk factors, risk conceptualization, risk scenarios, risk management strategies, and final judgment. In the present study, the risk assessment was based on records, so we were not able to carry out steps 3 to 7. In accordance with Brookstein [37], we assessed the presence of risk factors (step 2) by using a 3-point scale (present = 2, partially present = 1, not present = 0) and summed the scores of the 20 risk factors, which yielded a total score ranging from 0 to 40.

2.2.4 FAM

FAM was developed as an extension of HCR-20 v3 to predict the relapse risk for violence in women and considers eight additional items [27]. For the present study, the English version of the FAM was used. The manual states that the eight additional items should be rated on a five-point scale. Like the HCR-20 v3, the FAM also follows the structured professional judgement approach, meaning that ratings are not intended to be given numerical values. However, to ensure better comparability of findings in the present study, we rated the items by using an approach similar to that used for the HCR-20 v3, i.e., on a three-point scale (0 = no, 1 = possible or partial, 2 = yes), and summed the scores of the 10 FAM risk factors, which yielded a total score ranging from 0 to 20. The authors reported good interrater reliabilities for all FAM items, with intraclass correlation coefficients (ICCs) ranging from .63 to .97 [27].

Because the combined evaluation of the HCR-20 v3 and FAM included 18 items from the HCR-20 v3 and 20 items from the FAM, the total scores ranged from 0 to 56.

2.2.5 VRAG-R

VRAG-R [28] is an actuarial assessment tool used to predict violence relapses (German version by Rettenberger [38]). It rates twelve items with different scoring systems, as described in the VRAG-R manual. The predictive validity for violent relapses is reported to be good (AUC, .76) [38]. In accordance with the risk categories of the VRAG-R [38, page 5], we chose the cut-off value of 11, which distinguishes between individuals with recidivism rates below 45% and equal to or greater than 45% after 5 years and below 69% and equal to or greater than 69% after 12 years.

2.3 Procedures

Before the study, five research staff members (clinical psychologists) were trained in the prognostic instruments. Then, the staff members assessed patients by referring to the patient

medical records and coded the final risk judgements. To confirm a uniform standard of assessment ratings across reviewers, all five reviewers independently rated 11 patients with the five assessment tools and interrater reliabilities were calculated (the ICCs of the assessment tools were as follows: PCL-R, .71; LSI-R, .74; HCR-20 v3, .61; FAM, .82; and VRAG-R, .89). To evaluate actual relapses after patients had been discharged, we obtained information from the German Federal Central Criminal Register in September 2020 and February 2021, in which all formal convictions are documented. Each formal conviction documented after release from the hospital or prison (in the case of treatment discontinuation) was counted as recidivism. In addition, violent offenses, which were defined as convictions for an offense involving crimes against persons (e.g., homicide, sex crimes, assault, threat, and robbery), were analyzed separately. The time at risk began at the time of release from the forensic psychiatric hospital (or prison) and ended when another crime was committed. If no further crime was committed, the time at risk ended on the date when the report was obtained from the German Federal Central Criminal Register. The mean time at risk was 6.0 years (standard deviation [SD], 4.9 years).

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2013. All procedures involving patients were approved by the ethics committee of the Bavarian Medical Association, Germany (approval no.: 2019-167). Informed consent was not necessary because of the retrospective nature of the study. This was approved by the ethics committee.

2.4 Statistical analysis

Means, SDs, and absolute and relative frequencies were calculated to describe the sample. To test for group differences between recidivists and non-recidivists, we used the Mann-Whitney *U* test.

Predictive validity was determined by dichotomizing the assessment scores. We used the cut-off values recommended in the manuals (see Methods section), as follows: PCL-R, > 24; LSI-R, > 33; and VRAG-R, > 11. We included the following commonly used primary outcome values: sensitivity, specificity, positive predictive value, negative predictive value, *r* (point-biserial correlation coefficient), and the AUC statistics [39]. Sensitivity indicates how reliably the assessment instrument correctly predicted relapse in patients who relapsed, and specificity indicates how reliably it correctly predicted the absence of relapse in patients who did not relapse. Positive predictive values were defined as the proportion of patients classified as high risk who went on to (violently) reoffend, and negative predictive values, as the proportion of patients classified as low risk who did not go on to (violently) reoffend. Point-biserial correlation analysis is a statistical procedure for measuring and evaluating the strength and direction (i.e., positive versus negative correlation) of the relationship between two variables. It allows one to determine whether changes in one variable are associated with changes in another and quantifies the degree of this connection (strong effect, ± .50 and above; medium effect, between ± .30 and ± .49; small effect, ± .29 and below [40]). The AUC statistics were determined with a receiver operation characteristics (ROC) curve, which is the function of the rate of true positives (i.e., sensitivity) and the rate of false positives (i.e., 1-specificity). The AUC expresses the accuracy of a prognostic tool in discriminating between relapsed and non-relapsed patients. An AUC of .5 indicates chance-level accuracy. According to commonly accepted standards, AUC values greater than .714 are generally considered to indicate good prognostic instrument performance [41]. To determine whether the differences between the AUCs of the assessment tools were statistically significant, the Mann-Whitney *U* test was used.

Finally, we compared the predictive performance of the five instruments in two subgroups: patients with schizophrenia (n = 81) and patients with a substance use disorder (n = 393). To test for group differences between recidivists and non-recidivists, we used the

Mann-Whitney *U* test. In addition, we calculated the AUC statistics for both subgroups and compared the AUCs with the non-parametric Mann-Whitney *U* test.

2.5 Transparency and openness

We report how we determined our sample size, all data exclusions, and all measures in the study, and we follow JARS [42]. All data and analysis codes are available from the corresponding author, [JS], upon reasonable request. Data were analyzed with IBM SPSS Statistics for Windows version 26 (Armonk, NY: IBM Corp.). This study's design and its analysis were not pre-registered.

3 Results

3.1 Sociodemographic characteristics

Table 1 shows the detailed sociodemographic information of the sample and Table 2 lists the forensic psychiatric characteristics of the sample. All patients were female. The most common diagnoses were substance use disorder (*n* = 393) and schizophrenia (*n* = 81). About 15% of the patients had a comorbid personality disorder. 40% of the patients (*n* = 208) relapsed, 11% (*n* = 60) with a violent offense.

Table 1

Sociodemographic information of the patients (N =525)

	M (*SD*)
Age (at hospital admission)	34.15 (10.14)
	Frequency (%)
Marital status	
Single	302 (58%)
Married / In a registered partnership	73 (14%)
Widowed	15 (3%)

Separated / Divorced	135 (26%)
School and vocational training according to the International Standard Classification of Education	
No education	6 (1%)
Primary education	50 (10%)
Lower secondary education	233 (44%)
Upper secondary education	209 (40%)
Post-secondary non-tertiary education	10 (2%)
Tertiary education	16 (3%)
Occupation[a]	
Unemployed	410 (78%)
Employed	66 (13%)
Undergoing training	5 (1%)
Not capable of being employed	43 (8%)
Provision of parental care[b]	
No provision of parental care (despite having children)	128 (54%)
Sole parental caregiver	49 (21%)
Joint parental caregiver	61 (26%)

Note. [a]missing data = 1; [b]patients without children were not considered; *M*, mean; *SD*, standard deviation

Table 2

Forensic psychiatric characteristics of the patients (N = 525)

	M (*SD*)
Age at first crime (in years)[a]	23.94 (11.24)
Age at first inpatient treatment (in years)[b]	27.17 (10.73)
	Frequency (%)
Index offense	
Offense against public order	1 (.2%)
Sexual assault	1 (.2%)
Insult	1 (.2%)
Traffic offense	19 (4%)
Financial crime / Property damage	75 (14%)

Resistance against state authority	2 (.4%)
Coercion	9 (2%)
Robbery	29 (6%)
Drug-related crime	201 (38%)
Arson	29 (6%)
Assault	112 (21%)
Homicide	46 (9%)
Main clinical diagnosis and comorbid personality disorder[bc]	
F0: Organic disorder	4 (1%)
F10: Alcohol-related disorder	50 (10%)
F10 + F6: Alcohol-related disorder and personality disorder	21 (4%)
F11-18: Substance-related disorder to specific substance	118 (23%)
F11-18 + F6: Substance-related disorder to specific substance and personality disorder	10 (2%)
F19: Multiple drug use	157 (30%)
F19 + F6: Multiple drug use and personality disorder	37 (7%)
F2: Schizophrenic disorder	73 (14%)
F2 + F6: Schizophrenic disorder and personality disorder	8 (2%)
F3: Mood disorder	4 (.7%)
F3 + F6: Mood disorder and personality disorder	1 (.2%)
F4: Adjustment disorder / PTSD	1 (.2%)
F4 + F6: Adjustment disorder/PTSD and personality disorder	1 (.2%)
F6: Personality disorder	33 (6%)
F7: Mental retardation	1 (.2%)
F9: Conduct disorder	1 (.2%)
F9: Conduct disorder and personality disorder	2 (.4%)

Note. [a]missing data = 1; [b]missing data = 3; [c]diagnoses according to International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10), the diagnoses were made by physicians working in the forensic psychiatric hospital; PTSD, posttraumatic stress disorder; *M*, mean; *SD*, standard deviation

3.2 Predictive validity of assessment tools

Table 3 presents the metrics of the five predictive instruments regarding their prognostic validity for general recidivism. All instruments significantly distinguished between the group of patients with a recidivism offense and those without (see column test statistics). The confidence intervals of the AUC values for all five instruments encompassed the threshold of .714, indicating that none exceeded this value. Therefore, the prognostic validity can be interpreted as moderate to good. When comparing the AUC values of the instruments, the LSI-R proved to be the best instrument for predicting general recidivism in the present sample (LSI-R compared to PCL-R: $z = 2.956$, $p = .003$, $d = .260$; to HCR-20 v3: $z = 3.442$, $p = .001$, $d = .304$; to HCR-20 v3 and FAM: $z = 2.670$, $p = .008$, $d = .235$; and to VRAG-R: $z = 2.735$, $p = .006$, $d = .240$). Furthermore, the prediction of general recidivism with the combined HCR-20 v3 and FAM was not better than that with the HCR-20 v3 alone ($z = 1.780$, $p = .075$, $d = .156$). When considering the scales of the PCL-R, Factor 2 was a significantly better predictor of general recidivism than Factor 1 ($z = 4.298$, $p < .001$, $d = .382$). The HCR-20 v3 risk management subscale also performed significantly better than the HCR-20 v3 clinical scale ($z = 3.474$, $p = .001$, $d = .307$). All other pairwise comparisons of the AUC metrics not mentioned did not differ significantly from each other (AUC values at the item level can be found in Supplement 1).

Table 3

Means, standard deviations, test statistics of the Mann-Whitney U test comparing non-recidivists and recidivists, correlations, and area under the curve and confidence interval values of the assessment tools for general reoffending in the sample (N = 525)

	General reoffending			
	Non-recidivists	Recidivists	Test statistics	*r*, AUC, CI

	M (*SD*)	*M* (*SD*)		
PCL-R				
Total score	12.54 (7.14)	17.42 (6.97)	*z* = 7.364*** *d* = .678	*r* = .320*** AUC = .690*** CI = .644, .735
Factor 1	4.25 (3.52)	5.33 (3.49)	*z* = 3.588*** *d* = .317	*r* = .157*** AUC = .592*** CI = .543, .641
Factor 2	7.09 (4.37)	10.47 (4.52)	*z* = 8.032*** *d* = .749	*r* = .351*** AUC = .707*** CI = .661, .752
LSI-R	21.97 (7.97)	29.37 (7.44)	*z* = 9.639*** *d* = .927	*r* = .423*** AUC = .748*** CI = .707, .790
HCR-20 v3				
Total score	20.74 (7.09)	25.26 (6.21)	*z* = 7.102*** *d* = .651	*r* = .312*** AUC = .683*** CI = .637, .729
Historical	12.82 (3.28)	14.38 (2.60)	*z* = 5.545*** *d* = .499	*r* = .242*** AUC = .642*** CI = .595, .690
Clinical	2.84 (2.53)	3.98 (2.56)	*z* = 5.018*** *d* = .449	*r* = .219*** AUC = .628*** CI = .579, .677
Risk managment	5.08 (2.77)	6.90 (2.47)	*z* = 7.411*** *d* = .684	*r* = .324*** AUC = .690*** CI = .644, .738
HCR-20 v3 and FAM	26.98 (9.19)	33.44 (8.23)	*z* = 7.737*** *d* = .717	*r* = .338*** AUC = .699*** CI = .654, .744
VRAG-R	-4.40 (16.79)	7.30 (15.45)	*z* = 7.516*** *d* = .694	*r* = .332*** AUC = .694*** CI = .649, .739

Note. PCL-R, Psychopathy Checklist – Revised; LSI-R, Level of Service Inventory – Revised; HCR-20 v3, Historical Clinical Risk Management-20, version 3; FAM, Female Additional Manual; VRAG-R, Violence Risk Appraisal Guide – Revised; *M*, mean; *SD*, standard deviation; *z*, standardized test statistic of the Mann-Whitney *U* test; *d*, Cohen's *d* effect size for unequal sample sizes; *r*, point-biserial correlation; AUC, area under curve; CI, 95% confidence interval; ***$p < .001$, **$p < .01$ (two-tailed tested)

Table 4 show the sensitivities, specificities, positive and negative predictive values of the assessment tools PCL-R, LSI-R, and VRAG-R for general reoffending. All three instruments had very low sensitivity, which was due to the fact that only a few patients were correctly classified as positive (PCL-R, 14%; LSI-R, 35%; VRAG-R, 45%). The specificity was good, with a large proportion of patients correctly classified as negative (PCL-R, 94%; LSI-R, 91%; VRAG-R, 81%).

Table 4

Sensitivities, specificities, positive and negative predictive values of three assessment tools for general reoffending in the sample (N = 525)

	General reoffending			
	Sensitivity[1]	Specificity[2]	PPV[3]	NPV[4]
PCL-R	.14	.94	.63	63
LSI-R	.35	.91	.72	.68
VRAG-R	.45	.81	.60	.69

Note. PCL-R, Psychopathy Checklist – Revised, cut-off value > 24; LSI-R, Level of Service Inventory – Revised, cut-off value > 33; VRAG-R, Violence Risk Appraisal Guide – Revised, cut-off value > 11; [1]Sensitivity = Cp/(Cp + Fn); [2]Specificity = Cn/(Cn + Fp); [3]Positive predictive value = Cp/(Cp + Fp); [4]Negative predictive value = Cn/(Cn + Fn) (Cp, number of correct positive outcomes; Cn, number of correct negative outcomes; Fp, number of false positive outcomes; Fn number of false negative outcomes; PPV, positive predictive value; NPV, negative predictive value); as structured professional judgement tools, HCR-20 v3 and FAM do not specify cut-off scores for classifying assessed individuals into different risk levels. Therefore, specificity, sensitivity, positive predictive value, and negative predictive valuecould not be calculated.

Table 5 presents the metrics for predicting recidivism with a violent offense. Here, too, all methods can distinguish between recidivist and non-recidivist patients (see column test statistics). Again, the confidence intervals of the AUC values of all five prognostic instruments included the threshold of .714, so all instruments can be considered to be moderate to good. The comparison of the AUC metrics of the instruments showed the following differences: The VRAG-R predicted violent offenses less well than the HCR-20 v3

(z = 2.397, p = .017, d = .210) and the LSI-R (z = 2.074, p = .038, d = .182), and the HCR-20 v3 predicted violent offenses better than the HCR-20 v3 combined with the FAM (z = 2.504, p = .012, d = .220). All other pairwise comparisons of the AUC metrics between the instruments or the scales were not significant.

Table 5

Means, standard deviations, and test statistics of the Mann-Whitney U test comparing non-recidivists and recidivists and r, area under the curve, and confidence interval values of the assessment tools for violent reoffending in the sample (N = 525)

		Violent reoffending			
		Non-recidivists *M* (*SD*)	Recidivists *M* (*SD*)	Test statistics	*r*, AUC, CI
PCL-R					
	Total score	13.79 (7.22)	19.77 (7.24)	z = 5.654*** d = .509	r = .255*** AUC = .724*** CI = .656, .791
	Factor 1	4.47 (3.50)	6.28 (3.55)	z = 3.734*** d = .330	r = .163*** AUC = .647*** CI = .578, .717
	Factor 2	7.99 (4.52)	11.80 (4.95)	z = 5.663*** d = .510	r = .247*** AUC = .724*** CI = .650, .798
LSI-R		24.07 (8.32)	31.37 (7.66)	z = 6.096*** d = .552	r = .272*** AUC = .741*** CI = .675, .808
HCR-20 v3					
	Total score	21.80 (6.96)	28.17 (5.52)	z = 6.654*** d = .606	r = .285*** AUC = .764*** CI = .703, .824
	Historical	13.16 (3.08)	15.57 (2.51)	z = 5.870*** d = .530	r = .256*** AUC = .732*** CI = .666, .797
	Clinical	3.09 (2.57)	4.88 (2.34)	z = 5.114***	r = .223***

			d = .458	AUC = .701*** CI = .636, .766
	Risk managment	5.56 (2.77)	7.72 (2.18)	*z* = 5.847*** *d* = .528	*r* = .255*** AUC = .730*** CI = .667, .794
HCR-20 v3 and FAM		29.35 (9.72)	36.10 (7.88)	*z* = 5.766*** *d* = .520	*r* = .252*** AUC = .728*** CI = .665, .792
VRAG-R		-1.02 (17.00)	9.94 (15.99)	*z* = 4.571*** *d* = .407	*r* = .203*** AUC = .681*** CI = .611, .752

Note. PCL-R, Psychopathy Checklist – Revised; LSI-R, Level of Service Inventory – Revised; HCR-20 v3, Historical Clinical Risk Management-20, version 3; FAM, Female Additional Manual; VRAG-R, Violence Risk Appraisal Guide – Revised; *M*, mean; *SD*, standard deviation; *z*, standardized test statistic of the Mann-Whitney *U* test; *d*, Cohen's *d* effect size for unequal sample sizes; *r*, point-biserial correlation; AUC, area under curve; CI, 95% confidence interval, ***$p < .001$, **$p < .01$ (two-tailed tested)

Table 6 displays the sensitivities, specificities, and positive and negative predictive values of the assessment tools PCL-R, LSI-R, and VRAG-R for predicting violent reoffending. The sensitivity of the three instruments for predicting violent recidivism was very low (correctly classified as positive: PCL-R, 14%; LSI-R, 35%; VRAG-R, 45%). Overall, 143 of 525 (27%) patients recidivated with a violent offense without recidivism being predicted.

Table 6

Sensitivities, specificities, positive and negative predictive values of three assessment tools for violent reoffending in the sample (N = 525)

	Violent reoffending			
	Sensitivity[1]	Specificity[2]	PPV[3]	NPV[4]
PCL-R	.22	.92	.27	.90
LSI-R	.48	.85	.29	.93
VRAG-R	.50	.73	.19	.92

Note. PCL-R, Psychopathy Checklist – Revised, cut-off value > 24; LSI-R, Level of Service Inventory – Revised, cut-off value > 33; VRAG-R, Violence Risk Appraisal Guide – Revised, cut-off value > 11; [1]Sensitivity = Cp/(Cp + Fn); [2]Specificity = Cn/(Cn + Fp); [3]Positive predictive value = Cp/(Cp + Fp);

[4]Negative predictive value = Cn/(Cn + Fn) (Cp, number of correct positive outcomes; Cn, number of correct negative outcomes; Fp, number of false positive outcomes; Fn number of false negative outcomes; PPV, positive predictive value; NPV, negative predictive value); as structured professional judgement tools, HCR-20 v3 and FAM do not specify cut-off scores for classifying assessed individuals into different risk levels. Therefore, specificity, sensitivity, positive predictive value, and negative predictive value could not be calculated.

In a further analysis, we compared the predictive performance of the five instruments in two subgroups: patients with schizophrenia (*n*=81) and patients with a substance use disorder (*n*=393). In patients with schizophrenia, relapse occurred in 19% (*n* = 15) and was characterized by a violent offense in 11% (*n* = 9); in patients with substance use disorder, it occurred in 45% (*n* = 176) and was characterized by a violent offense in 12% (*n* = 46). Compared with patients with schizophrenia, patients with substance use disorder relapsed significantly more often with a general offense ($Chi^2(1) = 19.257$; $p < .010$; Cramer $V = .202$). Regarding violent recidivism, no differences were found between the two diagnostic groups ($Chi^2(1) = .023$; $p = .879$; Cramer $V = .007$). The results revealed no significant differences between the two groups in terms of the predictive validity of the five assessment tools; the AUC differences ranged from -.007 to .106 and did not differ from zero.

4 Discussion

The present study aimed to examine prognosis of recidivism in women treated in forensic psychiatric facilities by evaluating the predictive quality of common prognostic instruments (PCL-R, LSI-R, HCR-20 v3, HCR-20 v3 + FAM, and VRAG-R). After a mean observation period of 6.0 years (SD = 4.9 years), general recidivism had occurred in 208 (40%) of the 525 women examined, and violent recidivism in 60 (11%). These recidivism rates are similar to those observed in the studies by de Vogel et al. [21] and Schaap et al. [43]. De Vogel et al. [21] examined a sample of 71 women who were discharged from forensic psychiatric hospitals. After a mean follow-up period of 11.8 years (*SD* = 4.9), 24 (33.8%) were officially reconvicted for one or more offenses, and in 13 (18.3%) cases, these offenses were violent.

Schaap et al. [43] analyzed recidivism in 45 forensic inpatients and found that 16 (36%) were reconvicted of an offense (i.e., general recidivism) and 7 (16%) were reconvicted for a violent offense. We found not significant difference between these recidivism rates and those in the present sample (present study vs de Vogel et al.: $Chi^2(1) = 2.240$, $p = .135$, Cramer $V = .098$; present study vs Schaap et al.: $Chi^2(1) = .391$, $p = .532$, Cramer $V = .051$). For a comparison of the recidivism rates in the present female sample with a comparable male sample of forensic psychiatric inpatients, see Supplement 2.

The current study shows that the evaluated risk assessment tools are suitable for use in female forensic psychiatric patients because the tools were able to reliably differentiate between patients with and without general and violent recidivism. With regard to the quality of the predictions (AUC), the predictions were significantly better than chance, and the instruments consistently showed moderate to good performance. For general recidivism, the LSI-R had the best predictive quality, and for violent recidivism, the HCR-20 v3 and LSI-R both performed well. The present study further showed that supplementing the HCR-20 v3 with the FAM does not improve the prognosis, neither for the prediction of general recidivism nor for the prediction of violent recidivism. Thus, our data indicate that supplementary use of the FAM is not helpful in predicting recidivism.

When we compared the AUC values in our study with the prognostic validity of the risk assessment instruments HCR-20 v3, FAM, and PCL-R as reported by de Vogel et al. [21], we found similar, good metrics for predicting general recidivism (HCR-20 v3 = .667; HCR-20 v3 + FAM = .661 PCL-R = .601), but much better metrics in the present sample for predicting recidivism with a violent offense (HCR-20 v3 = .672; HCR-20 v3 + FAM = .651; PCL-R = .591). The better prognostic validity for recidivism with a violent offense in the present study may be because of the larger sample size. Generally, the AUC is not directly dependent on sample size, but it can indirectly be affected by it, especially when the sample is too small to contain a sufficient number of events needed for a reliable estimation of model

performance. For example, at a recidivism rate of 34% (the general recidivism rate found by de Vogel et al.), a total sample size of 60 people is sufficient to test a prognostic instrument with an AUC of .714 against the null hypothesis (AUC = .5). However, at a recidivism rate of 18% (the violent recidivism found by de Vogel et al.), a total sample size of 80 people would be needed [44].

In a further analysis, we examined whether predictive accuracy was influenced by patient diagnosis and found that it was not, i.e., all five prognostic instruments achieved comparably good results in both subgroups. Further differentiation of the instruments depending on individual diagnoses does not appear to be necessary when using the instruments in mentally ill women.

In the present study all prognostic instruments had rather low sensitivity for predicting general and violent offending. To improve sensitivity, the cut-off values could be decreased, although that usually results in a slight decrease in specificity. Tools with high sensitivity will be most effective at safeguarding the public and may also gain significant political support. However, tools with high specificity will best protect the rights and interests of psychiatric patients. Achieving a balance between false positives and false negatives is an ethical matter and depends on the social and political context in which the tool is being used. Therefore, the following recommendations can be derived for the use of these instruments in forensic psychiatric samples of women: The low sensitivity of the prognostic instruments means that they should not be used (solely) to make decisions about the timing of discharge from a forensic psychiatric hospital because they do not reliably classify patients who will relapse after discharge, so public safety may be compromised in some cases; however, the instruments can assist clinicians in developing risk management plans that can be used to reduce individual risk within the framework of therapeutic interventions or social-pedagogical support measures.

Future studies could examine various measures to enhance the sensitivity of instruments used in forensic psychiatric samples of women, such as altered cut-off scores, different weighting of individual risk factors, and the combined application of different instruments.

The present study has some limitations that should be considered when interpreting the results and drawing conclusions. First, this was a retrospective study in which the items of the prognostic instruments were rated with data collected from information in the files of patients who had already been discharged. As a result, missing data could not be ascertained retrospectively and the accuracy of the information could not be verified. Second, for the same reason, we were not able to apply the PCL-R and LSI-R in interview form. Third, the HCR-20 v3 is not designed for quantifying items, which limits the transferability of our results; however, the AUC value for predicting a violent offense based on the HCR-20 v3 total score was very good compared with that of the other instruments, suggesting that summing the risk factors yields excellent results, obviating the need to implement steps 3 to 7. Therefore, in routine clinical care, if professionals can only assess a patient based on records, they can achieve a good prognosis by summing the fulfilled factor values. Fourth, recidivism was assessed from entries in the Federal Central Register, so incidents from the dark figure of crime were not captured. Additionally, in accordance with Section 46 of the Federal Central Register Act (BZRG), entries in the Federal Central Register are deleted after the expiration of a specified time limit, which depends on the severity of the offense. Minor offenses (e.g., fines up to 90 daily rates) are deleted after 3 years, offenses resulting in a sentence of no more than one year of imprisonment are deleted after 5 years, and more serious offenses (e.g., sentences exceeding one year of imprisonment) are deleted after 10 years. And last, all instruments were scored by one scorer, so it is possible that the outcome of one instrument may have influenced the outcome of another.

5 Conclusions

A significant strength of the study lies in the large sample size (N = 525) and the extended observation period (mean time at risk, 6 years). Thus, the study yielded useful results on the prognostic validity and generalizability of the studied instruments. The AUC metrics indicate that all assessment instruments can be used to predict general and violent recidivism in women in forensic psychiatric care. In particular, the LSI-R appears to perform best in predicting general recidivism and both the HCR-20 v3 and LSI-R appear to perform equally well for the specific prediction of violent recidivism. All instruments exhibit low sensitivity and are not suitable as the sole basis for discharge decisions because they do not correctly classify a high proportion of patients who reoffend with a violent offense. Nevertheless, by highlighting individual risk areas, they can provide valuable information for planning therapy goals or support measures.

Funding

This work was supported by the Bavarian State Ministry for Family, Labor and Social Affairs, Germany [grant number ZBFS-X/1-10.700-5/3/9].

Acknowledgments

The codebook was designed in collaboration with the Office of Corrections and Rehabilitation, Zurich, Switzerland. The authors thank Jacquie Klesing, Board-certified Editor in the Life Sciences (ELS), for editing assistance with the manuscript.

Data statement

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

References

[1] Fazel S, Danesh J. Serious mental disorder in 23 000 prisoners: a systematic review of 62 surveys. The Lancet 2002;359:545–50. https://doi.org/10.1016/S0140-6736(02)07740-1.

[2] Fazel S, Hayes AJ, Bartellas K, Clerici M, Trestman R. Mental health of prisoners: prevalence, adverse outcomes, and interventions. Lancet Psychiatry 2016;3:871–81. https://doi.org/10.1016/S2215-0366(16)30142-0.

[3] Garofalo C, Velotti P. Negative emotionality and aggression in violent offenders: The moderating role of emotion dysregulation. J Crim Justice 2017;51:9–16. https://doi.org/10.1016/j.jcrimjus.2017.05.015.

[4] Newhill CE, Mulvey EP. Emotional dysregulation: The key to a treatment approach for violent mentally ill individuals. Clin Soc Work J 2002;30:157–71. https://doi.org/10.1023/A:1015293428307.

[5] Ullrich S, Keers R, Coid JW. Delusions, anger, and serious violence: new findings from the MacArthur Violence Risk Assessment Study. Schizophr Bull 2014;40:1174–81. https://doi.org/10.1093/schbul/sbt126.

[6] Ogilvie JM, Tzoumakis S, Thompson C, Allard T, Dennison S, Kisely S, et al. Psychiatric illness and the risk of reoffending: recurrent event analysis for an Australian birth cohort. BMC Psychiatry 2023;23:355. https://doi.org/10.1186/s12888-023-04839-0.

[7] Zgoba KM, Reeves R, Tamburello A, Debilio L. Criminal recidivism in inmates with mental illness and substance use disorders. J Am Acad Psychiatry Law 2020;48:209–15. https://doi.org/10.29158/JAAPL.003913-20.

[8] Okamura M, Okada T, Okumura Y. Recidivism among prisoners with severe mental disorders. Heliyon 2023;9. https://doi.org/10.1016/j.heliyon.2023.e17007.

[9] KiDeuk K, Becker-Cohen M, Serakos M. The Processing and Treatment of Mentally Ill Persons in the Criminal Justice System: A Scan of Practice and Background Analysis. Washington, DC: Urban Institute; 2015.

[10] Douglas T, Pugh J, Singh I, Savulescu J, Fazel S. Risk assessment tools in criminal justice and forensic psychiatry: The need for better data. Eur Psychiatry 2017;42:134–7. https://doi.org/10.1016/j.eurpsy.2016.12.009.

[11] Brown J, Singh JP. Forensic risk assessment: A beginner's guide. Arch Forensic Psychol 2014;1:49–59.

[12] Ogonah MG, Seyedsalehi A, Whiting D, Fazel S. Violence risk assessment instruments in forensic psychiatric populations: a systematic review and meta-analysis. Lancet Psychiatry 2023;10:780–9. https://doi.org/10.1016/S2215-0366(23)00256-0.

[13] Walmsley R. World female imprisonment list. Women and girls in penal institutions, including pre-trial detainees / remand prisoners. London: International centre for prison studies; 2017.

[14] Geraghty KA, Woodhams J. The predictive validity of risk assessment tools for female offenders: A systematic review. Aggress Violent Behav 2015;21:25–38. https://doi.org/10.1016/j.avb.2015.01.002.

[15] Katz RS. Explaining girl's and women's crime and desistance in the context of their victimization experiences – A developmental test of revised strain theory and the life course perspective. Violence Women 2000;6:633–60. https://doi.org/10.1177/1077801200006006005.
[16] van Voorhis P, Wright EM, Salisbury E, Bauman A. Women's risk factors and their contributions to existing risk/needs assessment. Crim Justice Behav 2010;37:261–88. https://doi.org/10.1177/0093854809357442.
[17] Johansson P, Kempf-Leonard K. A gender-specific pathway to serious, violent, and chronic offending? Exploring Howell's risk factors for serious delinquency. Crime Delinquency 2009;55:216–40. https://doi.org/10.1177/0011128708330652.
[18] Vogel V, Nicholls TL. Gender matters: An introduction to the special issues on women and girls. Int J Forensic Ment Health 2016;15:1–25. https://doi.org/10.1080/14999013.2016.1141439.
[19] Blanchett K, Brown SL. The assessment and treatment of women offenders: An integrative perspective. Chichester: John Wiley & Sons Ltd.; 2006.
[20] Nicholls TL, Brink J, Greaves C, Lussier P, Verdun-Jones S. Forensic psychiatric inpatients and aggression: an exploration of incidence, prevalence, severity, and interventions by gender. Int J Law Psychiatry 2009;32:23–30. https://doi.org/10.1016/j.ijlp.2008.11.007.
[21] Vogel V, Bruggeman M, Lancel M. Gender-sensitive violence risk assessment: Predictive validity of six tools in female forensic psychiatric patients. Crim Justice Behav 2019;46:528–49. https://doi.org/10.1177/0093854818824135.
[22] Müller JL, Saimeh N, Briken P, Eucker S, Hoffmann K, Koller M, et al. Standards für die Behandlung im Maßregelvollzug nach §§ 63 und 64 StGB. Forensische Psychiatr Psychol Kriminol 2018;12:93–125. https://doi.org/10.1007/s11757-017-0445-0.
[23] Mayer J, Wolf V, Steiner I, Franke I, Klein V, Streb J, et al. Rückfallprognose bei Straftäterinnen. Forensische Psychiatr Psychol Kriminol 2023;17:189–98. https://doi.org/10.1007/s11757-023-00770-y.
[24] Hare RD. Hare Psychopathy Checklist-Revised (PCL-R). Toronto: Multi-Health Systems; 2003.
[25] Andrews D, Bonta J. The Level of Service Inventory-Revised (LSI-R). Toronto: Multi Health Systems; 1995.
[26] Douglas KS, Hart SD, Webster CD, Belfrage H. HCR-20 V3: Assessing risk for violence – User guide. Simon Fraser University, Canada: Mental Health, Law, and Policy Institute; 2013.
[27] Vogel V, Vries Robbé M, van Kalmthout W, Place C. Female Additional Manual (FAM): Additional guidelines to the HCR-20 V3 for assessing risk for violence in women. Utrecht: Van der Hoeven Kliniek; 2014.
[28] Rice ME, Harris GT, Lang C. Validation of and revision to the VRAG and SORAG: the Violence Risk Appraisal Guide-Revised (VRAG-R). Psychol Assess 2013;25:951–65. https://doi.org/10.1037/a0032878.
[29] Mokros A, Hollerbach P, Nitschke J, Habermeyer E. PCL-R. Hare Psychopathy Checklist – Revised. Deutsche Version der Hare Psychopathy Checklist – Revised (PCL-R) von R. D. Hare. Göttingen: Hogrefe; 2017.
[30] Cooke DJ, Michie C. An item response theory evaluation of Hare's Psychopathy Checklist. Psychol Assess 1997;9:2–13. https://doi.org/10.1037/1040-3590.9.1.3.
[31] Hare RD, Harpur TJ, Hakstian AR, Forth AE, Hart SD, Newman JP. The revised psychopathy checklist: reliability and factor structure. Psychol Assess J Consult Clin Psychol 1990;2:338. https://doi.org/10.1037/1040-3590.2.3.338.

[32] Harris GT, Rice ME, Quinsey VL. Psychopathy as a taxon: evidence that psychopaths are a discrete class. J Consult Clin Psychol 1994;62:387. https://doi.org/10.1037/0022-006X.62.2.387.
[33] Dahle K-P, Schneider V, Ziethen F. Standardisierte Instrumente zur Kriminalprognose. Forensische Psychiatr Psychol Kriminol 2007;1:15–26. https://doi.org/10.1007/s11757-006-0004-6.
[34] Dahle K-P, Harwardt F, Schneider-Njepel V. Inventar zur Einschätzung des Rückfallrisikos und des Betreuungs- und Behandlungsbedarfs von Straftätern: LSI-R; deutsche Version des Level of Service Inventory-Revised nach Don Andrews und James Bonta. Göttingen: Hogrefe; 2012.
[35] Bolzmacher M, Born P, Eucker S, Franqué F, Holzinger B, Kötter S, et al. Die Vorhersage von Gewalttaten mit dem HCR-20 v3. Gießen: Institut für forensische Psychiatrie Haina e. V.; 2014.
[36] Franqué F. HCR-20 - Historical-Clinical-Risk Management-20 Violence Risk Assessment Scheme. In: Rettenberger M, Franqué F, editors. Handb. Kriminalprognostischer Verfahr., Göttingen: Hogrefe; 2013, p. 141–58.
[37] Brookstein DM, Daffern M, Ogloff JRP, Campbell RE, Chu CM. Predictive validity of the HCR-20 ^{V3} in a sample of Australian forensic psychiatric patients. Psychiatry Psychol Law 2021;28:325–42. https://doi.org/10.1080/13218719.2020.1775152.
[38] Rettenberger M, Gregório Hertz P, Eher R. Die deutsche Version des Violence Risk Appraisal Guide-Revised (VRAG-R). Wiesbaden: Kriminologische Zentralstelle; 2017.
[39] Rice ME, Harris GT. Violent recidivism: Assessing predictive validity. J Consult Clin Psychol 1995;63:737–48. https://doi.org/10.1037//0022-006x.63.5.737.
[40] Cohen J. Statistical power analysis for the behavioral sciences. routledge; 2013.
[41] Rice ME, Harris GT. Comparing Effect Sizes in Follow-Up Studies: ROC Area, Cohen's d, and r. Law Hum Behav 2005;29:615–20. https://doi.org/10.1007/s10979-005-6832-7.
[42] Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. Am Psychol 2018;73:3. https://psycnet.apa.org/doi/10.1037/amp0000389.
[43] Schaap G, Lammers S, Vogel V. Risk assessment in female forensic psychiatric patients: A quasi-prospective study into the validity of the HCR-20 and PCL-R. J Forensic Psychiatry Psychol 2009;20:354–65. https://doi.org/10.1080/14789940802542873.
[44] Lu G. Sample size formulas for estimating areas under the receiver operating characteristic curves with precision and assurance. Master's Thesis. The University of Western Ontario (Canada), 2021.

Supplement 1.

Table 1

Predicitive validity of the individual items

	General reoffending			Violent reoffending		
	AUC	95%-Confidence interval		AUC	95%-Confidence interval	
PCL-R Item 1	.518	.467	.569	.546	.463	.628
PCL-R Item 2	.509	.458	.560	.523	.443	.602
PCL-R Item 3	.575	.525	.625	.577	.497	.656
PCL-R Item 4	.572	.522	.623	.525	.444	.605
PCL-R Item 5	.611	.562	.660	.564	.488	.641
PCL-R Item 6	.555	.505	.606	.637	.558	.716
PCL-R Item 7	.495	.444	.545	.553	.473	.632
PCL-R Item 8	.509	.459	.560	.574	.492	.656
PCL-R Item 9	.574	.524	.623	.567	.493	.640
PCL-R Item 10	.553	.503	.603	.640	.566	.713
PCL-R Item 11	.553	.502	.604	.578	.498	.658
PCL-R Item 12	.565	.514	.616	.594	.512	.676
PCL-R Item 13	.590	.541	.640	.640	.572	.709
PCL-R Item 14	.617	.568	.666	.626	.552	.700
PCL-R Item 15	.650	.602	.698	.631	.556	.706
PCL-R Item 16	.553	.503	.603	.601	.528	.674
PCL-R Item 17	.578	.527	.628	.534	.455	.614
PCL-R Item 18	.599	.548	.649	.630	.550	.710
PCL-R Item 19	.655	.608	.703	.592	.518	.666
PCL-R Item 20	.639	.590	.688	.625	.546	.704
LSI-R Item 1	.584	.527	.641	.483	.394	.572
LSI-R Item 2	.621	.564	.677	.517	.429	.604
LSI-R Item 3	.627	.570	.684	.545	.457	.633
LSI-R Item 4	.524	.465	.583	.436	.348	.524
LSI-R Item 5	.544	.485	.604	.539	.448	.630

LSI-R Item 6	.500	.441	.559	.500	.412	.588
LSI-R Item 7	.503	.444	.562	.508	.419	.598
LSI-R Item 8	.571	.513	.629	.587	.503	.671
LSI-R Item 9	.633	.577	.689	.569	.484	.654
LSI-R Item 10	.540	.481	.598	.645	.567	.722
LSI-R Item 11	.608	.551	.665	.631	.552	.710
LSI-R Item 12	.615	.558	.671	.626	.548	.704
LSI-R Item 13	.575	.516	.634	.638	.551	.725
LSI-R Item 14	.512	.453	.571	.461	.373	.549
LSI-R Item 15	.536	.476	.595	.587	.494	.679
LSI-R Item 16	.531	.473	.590	.511	.424	.599
LSI-R Item 17	.545	.486	.605	.541	.450	.632
LSI-R Item 18	.612	.555	.669	.627	.548	.706
LSI-R Item 19	.602	.545	.659	.612	.532	.692
LSI-R Item 20	.608	.551	.665	.641	.565	.717
LSI-R Item 21	.631	.574	.687	.636	.557	.715
LSI-R Item 22	.556	.498	.614	.588	.508	.667
LSI-R Item 23	.632	.575	.689	.593	.508	.679
LSI-R Item 24	.563	.505	.621	.639	.562	.716
LSI-R Item 25	.574	.516	.632	.612	.530	.695
LSI-R Item 26	.599	.540	.657	.530	.441	.619
LSI-R Item 27	.574	.515	.633	.582	.491	.673
LSI-R Item 28	.526	.466	.585	.513	.423	.602
LSI-R Item 29	.507	.447	.566	.509	.420	.598
LSI-R Item 30	.580	.523	.638	.589	.509	.668
LSI-R Item 31	.590	.533	.647	.614	.539	.688
LSI-R Item 32	.488	.429	.547	.580	.493	.668
LSI-R Item 33	.609	.553	.665	.556	.472	.640
LSI-R Item 34	.621	.564	.679	.571	.484	.659
LSI-R Item 35	.577	.518	.636	.557	.466	.647
LSI-R Item 36	.580	.521	.639	.540	.450	.631
LSI-R Item 37	.587	.529	.644	.624	.544	.703

LSI-R Item 38	.555	.497	.613	.474	.385	.564
LSI-R Item 39	.589	.531	.648	.689	.607	.771
LSI-R Item 40	.563	.506	.621	.482	.393	.571
LSI-R Item 41	.591	.533	.650	.633	.544	.721
LSI-R Item 42	.578	.518	.637	.565	.473	.657
LSI-R Item 43	.591	.532	.650	.613	.522	.705
LSI-R Item 44	.540	.481	.600	.562	.469	.656
LSI-R Item 45	.567	.508	.627	.614	.521	.707
LSI-R Item 46	.567	.510	.625	.573	.492	.654
LSI-R Item 47	.497	.438	.556	.497	.409	.584
LSI-R Item 48	.532	.474	.591	.516	.429	.603
LSI-R Item 49	.500	.441	.559	.500	.412	.588
LSI-R Item 50	.522	.462	.581	.546	.453	.639
LSI-R Item 51	.605	.546	.663	.651	.564	.738
LSI-R Item 52	.557	.497	.617	.560	.467	.652
LSI-R Item 53	.559	.499	.618	.596	.503	.688
LSI-R Item 54	.607	.548	.666	.599	.508	.689
HCR-20 v3 Item H1	.516	.466	.567	.639	.571	.706
HCR-20 v3 Item H2	.553	.504	.603	.541	.468	.614
HCR-20 v3 Item H3	.548	.499	.598	.528	.453	.603
HCR-20 v3 Item H4	.574	.525	.623	.589	.520	.657
HCR-20 v3 Item H5	.562	.513	.612	.522	.446	.597
HCR-20 v3 Item H6	.453	.403	.503	.473	.396	.549
HCR-20 v3 Item H7	.570	.520	.620	.617	.540	.694
HCR-20 v3 Item H8	.563	.514	.612	.549	.477	.622
HCR-20 v3 Item H9	.542	.491	.593	.654	.573	.736
HCR-20 v3 Item H10	.598	.549	.646	.611	.546	.676
HCR-20 v3 Item C1	.607	.557	.657	.672	.603	.740
HCR-20 v3 Item C2	.517	.466	.568	.538	.456	.619
HCR-20 v3 Item C3	.503	.452	.554	.541	.461	.622
HCR-20 v3 Item C4	.623	.574	.672	.673	.605	.741
HCR-20 v3 Item C5	.609	.559	.660	.643	.568	.718

HCR-20 v3 Item R1	.644	.595	.693	.675	.604	.746
HCR-20 v3 Item R2	.617	.569	.665	.634	.564	.704
HCR-20 v3 Item R3	.596	.547	.645	.652	.586	.719
HCR-20 v3 Item R4	.633	.584	.681	.674	.610	.737
HCR-20 v3 Item R5	.635	.587	.683	.626	.557	.695
FAM Item H7	.611	.561	.661	.666	.594	.738
FAM Item H8	.565	.515	.614	.551	.478	.623
FAM Item H11	.507	.456	.558	.473	.397	.549
FAM Item H12	.565	.514	.616	.466	.386	.545
FAM Item H13	.565	.514	.616	.534	.454	.614
FAM Item H14	.524	.473	.575	.547	.470	.625
FAM Item C6	.633	.584	.682	.613	.538	.687
FAM Item C7	.547	.496	.598	.525	.441	.609
FAM Item R6	.581	.530	.632	.499	.416	.581
FAM Item R7	.630	.582	.679	.604	.531	.677
VRAG-R Item 1	.532	.481	.583	.530	.451	.608
VRAG-R Item 2	.576	.525	.626	.625	.548	.703
VRAG-R Item 3	.640	.591	.689	.602	.522	.683
VRAG-R Item 4	.509	.458	.560	.535	.454	.617
VRAG-R Item 5	.651	.604	.698	.568	.493	.643
VRAG-R Item 6	.648	.601	.696	.577	.501	.652
VRAG-R Item 7	.591	.541	.640	.609	.535	.683
VRAG-R Item 8	.559	.508	.610	.632	.551	.713
VRAG-R Item 9	.626	.576	.675	.600	.519	.680
VRAG-R Item 10	.629	.580	.678	.619	.534	.705
VRAG-R Item 11	.498	.447	.549	.499	.420	.578
VRAG-R Item 12	.664	.633	.725	.676	.639	.774

Table 2

Point-biserial correlation coefficients of the individual items

	General reoffending	Violent reoffending

	r	*p*	*r*	*p*
PCL-R Item 1	.045	.304	.072	.102
PCL-R Item 2	.029	.503	.051	.245
PCL-R Item 3	.144	.001	.099	.023
PCL-R Item 4	.150	.001	.030	.498
PCL-R Item 5	.201	<.001	.070	.107
PCL-R Item 6	.108	.014	.174	<.001
PCL-R Item 7	-.008	.851	.075	.086
PCL-R Item 8	.018	.684	.108	.013
PCL-R Item 9	.134	.002	.083	.058
PCL-R Item 10	.098	.024	.171	<.001
PCL-R Item 11	.112	.010	.104	.017
PCL-R Item 12	.133	.002	.122	.005
PCL-R Item 13	.166	<.001	.170	<.001
PCL-R Item 14	.212	<.001	.143	.001
PCL-R Item 15	.268	<.001	.147	.001
PCL-R Item 16	.100	.021	.124	.004
PCL-R Item 17	.141	.001	.035	.417
PCL-R Item 18	.199	<.001	.166	<.001
PCL-R Item 19	.301	<.001	.119	.006
PCL-R Item 20	.265	<.001	.150	.001
LSI-R Item 1	.203	<.001	-.016	.717
LSI-R Item 2	.259	<.001	.027	.536
LSI-R Item 3	.260	<.001	.051	.243
LSI-R Item 4	.081	.064	-.068	.118
LSI-R Item 5	.089	.041	.097	.026
LSI-R Item 7	.016	.719	.048	.273
LSI-R Item 8	.149	.001	.146	.001
LSI-R Item 9	.291	<.001	.101	.020
LSI-R Item 10	.034	.436	.182	<.001
LSI-R Item 11	.181	<.001	.124	.004
LSI-R Item 12	.218	<.001	.163	<.001

LSI-R Item 13	.142	.001	.188	<.001
LSI-R Item 14	.037	.442	-.024	.621
LSI-R Item 15	.069	.116	.098	.024
LSI-R Item 16	.051	.240	.005	.901
LSI-R Item 17	.095	.030	.092	.037
LSI-R Item 18	.210	<.001	.141	.001
LSI-R Item 19	.199	<.001	.131	.003
LSI-R Item 20	.210	<.001	.165	<.001
LSI-R Item 21	.239	<.001	.163	<.001
LSI-R Item 22	.098	.025	.082	.062
LSI-R Item 23	.239	<.001	.110	.012
LSI-R Item 24	.086	.048	.151	.001
LSI-R Item 25	.133	.002	.094	.033
LSI-R Item 26	.248	<.001	.046	.290
LSI-R Item 27	.155	<.001	.120	.006
LSI-R Item 28	.085	.051	.048	.273
LSI-R Item 29	.001	.982	.027	.535
LSI-R Item 30	.193	<.001	.144	.001
LSI-R Item 31	.224	<.001	.169	<.001
LSI-R Item 32	-.033	.445	.098	.025
LSI-R Item 33	.255	<.001	.109	.013
LSI-R Item 34	.269	<.001	.122	.005
LSI-R Item 35	.230	<.001	.098	.026
LSI-R Item 36	.256	<.001	.100	.023
LSI-R Item 37	.135	.002	.159	<.001
LSI-R Item 38	.164	<.001	-.031	.473
LSI-R Item 39	.170	<.001	.250	<.001
LSI-R Item 40	.169	<.001	-.028	.515
LSI-R Item 41	.217	<.001	.152	<.001
LSI-R Item 42	.243	<.001	.072	.102
LSI-R Item 43	.229	<.001	.131	.003
LSI-R Item 44	.137	.002	.093	.033

LSI-R Item 45	.190	<.001	.147	.001
LSI-R Item 46	.115	.009	.076	.083
LSI-R Item 47	-.076	.082	-.031	.478
LSI-R Item 48	.073	.095	.055	.206
LSI-R Item 50	.068	.121	.128	.003
LSI-R Item 51	.183	<.001	.208	<.001
LSI-R Item 52	.182	<.001	.161	<.001
LSI-R Item 53	.099	.023	.154	<.001
LSI-R Item 54	.232	<.001	.146	.001
HCR-20 v3 Item H1	.033	.446	.179	<.001
HCR-20 v3 Item H2	.162	<.001	.080	.067
HCR-20 v3 Item H3	.155	<.001	.059	.179
HCR-20 v3 Item H4	.178	<.001	.136	.002
HCR-20 v3 Item H5	.184	<.001	.041	.345
HCR-20 v3 Item H6	-.088	.045	-.031	.472
HCR-20 v3 Item H7	.128	.003	.144	.001
HCR-20 v3 Item H8	.162	<.001	.082	.061
HCR-20 v3 Item H9	.107	.014	.256	<.001
HCR-20 v3 Item H10	.210	<.001	.155	<.001
HCR-20 v3 Item C1	.197	<.001	.205	<.001
HCR-20 v3 Item C2	.092	.035	.136	.002
HCR-20 v3 Item C3	.007	.875	.057	.195
HCR-20 v3 Item C4	.219	<.001	.204	<.001
HCR-20 v3 Item C5	.200	<.001	.170	<.001
HCR-20 v3 Item R1	.278	<.001	.220	<.001
HCR-20 v3 Item R2	.225	<.001	.165	<.001
HCR-20 v3 Item R3	.178	<.001	.184	<.001
HCR-20 v3 Item R4	.239	<.001	.204	<.001
HCR-20 v3 Item R5	.265	<.001	.159	<.001
FAM Item H7	.191	<.001	.194	<.001
FAM Item H8	.164	<.001	.083	.057
FAM Item H11	.021	.639	-.051	.249

FAM Item H12	.120	.006	-.035	.428
FAM Item H13	.121	.005	.050	.256
FAM Item H14	.046	.290	.061	.166
FAM Item C6	.239	<.001	.136	.002
FAM Item C7	.089	.043	.032	.458
FAM Item R6	.149	.001	.005	.903
FAM Item R7	.240	<.001	.128	.003
VRAG-R Item 1	.060	.167	.050	.251
VRAG-R Item 2	.137	.002	.140	.001
VRAG-R Item 3	.246	<.001	.117	.007
VRAG-R Item 4	.037	.395	.064	.140
VRAG-R Item 5	.272	<.001	.091	.037
VRAG-R Item 6	.295	<.001	.108	.013
VRAG-R Item 7	.159	<.001	.112	.010
VRAG-R Item 8	.122	.005	.182	<.001
VRAG-R Item 9	.231	<.001	.119	.006
VRAG-R Item 10	.232	<.001	.135	.002
VRAG-R Item 11	-.035	.418	-.016	.720
VRAG-R Item 12	.286	<.001	.205	<.001

Supplement 2

Comparison with a male sample

Relapse rate can be influenced by many factors. For example, the length of the time at risk is important in that the longer the time at risk, the higher the relapse rate. In addition, relapse rate can also depend on the composition of the sample, and the literature provides evidence that patients with a substance use disorder are particularly prone to relapse. Both these factors are relevant in the present sample. Therefore, for our comparison with male patients, we chose a sample that had a similar time at risk and a similar prevalence of diagnoses as the female patients in our study.

There is evidence in the literature that women are less likely to reoffend than men [1,2]. However, the present study does not support this finding because rates were similar to those seen in men: Hogan and Olver [3] analyzed recidivism rates in 82 forensic psychiatric patients (93.3% male) over a mean period of 8.2 years and found general recidivism in 28% and violent recidivism in 17.1%; these rates do not differ significantly from those observed in the present female sample (general recidivism: $Chi^2(1) = 3.79$, $p = .052$; Cramer $V = .084$; violent recidivism: $Chi^2(1) = 1.92$, $p = .166$; Cramer $V = .064$).

Thus, our data suggest that recidivism rates are not lower in female than in male forensic psychiatric patients. It is possible that the effect described in the literature (see 7,21) only applies to mentally healthy female offenders.

References

[1] Emeka TQ, Sorensen JR. Female juvenile risk – Is there a need for gendered assessment instruments? Youth Violence and Juvenile Justice 2009;7:313–30.
[2] Vogel V, Bouman YHA, Horst P, Stam J, Lancel M. Gewalttätige Frauen: eine Multicenter-Studie über Genderunterschiede in der forensischen Psychiatrie. Forensische Psychiatrie Und Psychotherapie 2016;23:279–302.
[3] Hogan NR, Olver ME. Static and Dynamic Assessment of Violence Risk Among Discharged Forensic Patients. Crim Justice Behav 2019;46:923–38. https://doi.org/10.1177/0093854819846526.