Development and Preliminary Validation of a Fall Risk Assessment Tool for Psychiatric Inpatients: An Exploratory Sequential Mixed-Methods Study

Xiuqiong Gao*, Qiongxian Zhao, Huifang Li, Chengbao Zhu, Feng Guo, Wenhui Xi

Chuxiong Prefecture Second People's Hospital, Chuxiong, Yunnan, China

Abstract: Falls are a major patient-safety concern in psychiatric inpatient settings, where conventional risk factors intersect with psychiatric symptoms, adverse effects of psychotropic medication, fluctuating cognition, and urgent toileting needs. We used an exploratory sequential mixed-methods design to develop and preliminarily validate a fall risk assessment tool for adult psychiatric inpatients. Interviews with 10 psychiatric nurses and five inpatients informed the initial item pool. The tool was then refined through two Delphi rounds involving 15 experts and weighted with a group-consensus analytic hierarchy process (AHP) matrix. Preliminary validation assessed inter-rater agreement, content validity, convergence with the Morse Fall Scale, and exploratory 1-month risk discrimination. The final tool comprised nine risk domains and 36 ordered scoring descriptors. Both Delphi rounds had 100% response rates. In round 2, absolute endorsement was high across the scoring cells (mean range, 4.53–5.00; CV range, .00–.14), although the tie-corrected 45-cell Friedman analysis showed statistically detectable but weak rank-order concordance (Kendall's W = .106, $\chi^2(44) = 70.06$, $p = .007$), consistent with ceiling effects and numerous tied ratings. The group AHP matrix showed acceptable consistency (CR = .024), and overall content validity was acceptable (S-CVI/Ave = .92). Among 20 inter-rater cases, ICC(2,1) was .986, risk-category agreement was 80.0%, and linearly weighted κ was .762. Among 80 convergent-validity cases, correlations with the Morse Fall Scale were strong (Pearson $r = .848$; Spearman $\rho = .777$). In the 50-patient prospective sample, the AUC was .876 (95% CI, .673–1.000). The Youden cut point was ≥12; at the exploratory clinical high-risk cut point of ≥14, sensitivity was .600 and specificity was .800. The tool demonstrated preliminary content relevance, inter-rater agreement, and convergence with the Morse Fall Scale. Its risk-discrimination performance requires confirmation in a larger multicentre external validation study.

***Keywords*:** psychiatric nursing; inpatient falls; fall risk assessment; Delphi method; analytic hierarchy process; mixed methods; preliminary validation

1. Introduction

Falls are a persistent safety problem in psychiatric inpatient care. Although general fall-risk factors such as advanced age, previous falls, impaired gait, comorbid physical illness, and medication exposure are relevant across hospital settings, psychiatric wards have additional contextual risks. Antipsychotic-related adverse effects, orthostatic hypotension, excessive sedation, command hallucinations, persecutory delusions, impulsivity, fluctuating insight, confusion, and urgent toileting needs can rapidly change a patient's immediate risk state (Rao et al., 2018; Siafis et al., 2023; World Health Organization [WHO], 2021). Clinical guidance emphasizes systematic fall prevention, but the direct transfer of tools designed for general medical wards may miss risk situations that are specific to psychiatric inpatient care (Huang et al., 2023; Shi & Zhu, 2015).

Previous work has attempted to refine psychiatric fall-risk instruments. For example, Shen et al. (2021) simplified and tested a psychiatric fall-risk scale using 1-year admission data and reported psychometric and receiver operating characteristic (ROC) results. However, the applicability of existing instruments varies by setting, patient population, and scoring structure. In China, psychiatric inpatient care commonly occurs in closed or semi-closed wards where behavioral symptoms, psychotropic medication regimens, and safety-management practices interact with conventional fall-risk factors (Chen et al., 2023; Duan et al., 2026). Mental health risk assessment practices are not standardized across settings, and the choice and implementation of structured tools require attention to evidence, clinical context, and feasibility (Janse van Rensburg & van der Wath, 2020).

We treated the tool as a formative clinical risk assessment instrument rather than a reflective psychological scale measuring a single latent construct (Bollen & Lennox, 1991). The risk domains therefore need not be highly intercorrelated, and their AHP weights represent expert judgments about relative importance rather than empirically estimated contributions to fall outcomes. Given the exploratory validation sample and small number of falls, we aimed to develop the tool and obtain

initial evidence on content validity, feasibility, inter-rater agreement, convergent validity, and short-term risk discrimination, without claiming definitive predictive validity.

2. Methods

2.1 Design

An exploratory sequential mixed-methods design was used. The first phase used qualitative interviews to identify psychiatric-specific fall-risk themes from nurses' and patients' experiences. The second phase integrated qualitative findings and literature evidence to construct the tool through Delphi consultation and analytic hierarchy process (AHP) weighting. The third phase conducted preliminary clinical validation, including inter-rater agreement, content validity, convergent validity with the Morse Fall Scale, and an exploratory 1-month fall outcome analysis.

The study was conducted from July 1, 2025, to June 30, 2026. The Ethics Committee of Chuxiong Prefecture Second People's Hospital approved the study on April 18, 2025 (Approval No. 2025-07-Ke). All participating patients and nurses provided written informed consent.

2.2 Research Team

A five-member research team was established, including one chief nurse, one associate chief nurse, and three senior nurses. All members had more than 8 years of psychiatric clinical nursing or nursing-management experience. The team was responsible for study design and quality control, expert selection and Delphi coordination, qualitative interviews and coding, data management, and statistical analysis.

2.3 Phase 1: Qualitative Interviews

Purposive maximum-variation sampling was used to recruit 15 participants: 10 psychiatric nurses with at least 5 years of clinical experience and direct experience managing inpatient fall events, and five clinically stable inpatients who had experienced a fall during the preceding year. Before each interview, informed consent was obtained. A semi-structured interview guide was developed from the literature and reviewed by two psychiatric nursing experts. Interviews were audio-recorded and transcribed within 24 hours. Two researchers independently coded the transcripts, and disagreements were resolved through discussion with a third researcher. No new substantive theme emerged after the 13th interview; two additional interviews were conducted to confirm data saturation.

Four psychiatric-specific risk themes were identified: adverse effects of antipsychotic medication, symptom-related risk, altered consciousness and cognition, and impaired toileting or excretion ability. These themes were used to supplement conventional fall-risk factors and informed four psychiatric-specific risk domains in the final tool (Table 1).

Table 1. Qualitative Themes, Representative Quotations, and Translation Into Risk Domains

Qualitative theme	Representative quotation	Corresponding risk domain	Basis for scoring translation
Adverse effects of antipsychotic medication	"After taking clozapine, some patients stand up and sway, saying that their feet feel as if they are stepping on cotton. One older patient got up to use the toilet at night after medication and knelt down beside the bed because his legs became weak." — Nurse N03	Adverse effects of antipsychotic medication	Scored from 0 to 3 according to influence on activity: none; present but not affecting activity; clearly affecting activity; causing activity limitation or requiring intervention.
Symptom-related risk	"A patient said there was a voice in his head repeatedly telling him to stand by the window. He could not control himself and walked over; only after he became clear did he realize he had almost climbed out." — Nurse N07	Symptom-related risk	Scored from 0 to 3 according to predictability and controllability of symptom-driven risk behavior: no symptoms; distractibility; predictable risky behavior; uncontrollable risky behavior.
Altered consciousness and cognition	"She was fine during the day, but around dusk she became confused and could not recognize people. At night she insisted on going to the toilet; we could not stop her, and she fell in a hurry." — Nurse N05	Altered consciousness and cognition	Scored from 0 to 3 according to consciousness and cooperation: alert and cooperative; mild drowsiness or sedation; confusion or stupor; delirium.
Impaired toileting or excretion ability	"I need to get up four or five times a night. I cannot hold urine and walk quickly. The corridor light was dim, and once I almost slipped at the bathroom door." — Patient P02	Toileting/ excretion ability	Scored from 0 to 3 according to influence on toileting behavior: independent without nocturia; urinary frequency/nocturnal toileting; incontinence or requiring assistance; urgent incontinence or nocturnal toileting ≥3 times.

Note: Representative quotations were taken from de-identified interview transcripts. N = nurse; P = patient.

2.4 Phase 2: Tool Development and Delphi Consultation

The initial item pool was constructed from the four qualitative themes, the Morse Fall Scale framework, and the literature. General risk factors included previous falls, comorbid physical disease, transfer and mobility, age, and high-risk medication. The first draft included 11 risk domains and 42 scoring descriptors. After review by three senior psychiatric nursing experts, semantically overlapping or operationally unclear content was merged or revised, yielding a draft of nine risk domains and 37 scoring descriptors. The final 36 descriptors were mutually exclusive ordered scoring descriptions within the nine domains, rather than 36 independent reflective measurement items.

The Delphi questionnaire included study background, Likert-type 5-point importance ratings for risk domains and scoring descriptors, open-ended revision suggestions, and expert self-ratings of authority. Fifteen experts from tertiary psychiatric hospitals or psychiatric departments of general hospitals in China were invited. Inclusion criteria were at least 10 years of psychiatric clinical, nursing-management, or psychiatric medical experience; a bachelor's degree or higher; an intermediate professional title or higher; and familiarity with fall prevention. The expert panel covered psychiatric clinical nursing, nursing management, psychiatry, clinical pharmacy, and rehabilitation medicine. The same 15 experts completed both Delphi rounds.

Anonymous consultation was conducted online. In round 1, 15 experts evaluated 37 scoring descriptors and provided revision suggestions. The research team deleted or revised content based on importance mean, coefficient of variation (CV), and qualitative comments. A minimum mean importance score of 4.50 and CV lower than .25 were used as criteria for retention (Li & Chen, 2018; Liu et al., 2020). The second-round questionnaire included nine domains and 36 scoring descriptors. Both rounds achieved a 100% valid response rate.

Analytic Hierarchy Process Weighting and Scoring: The two-level AHP model comprised the target, "fall risk assessment for psychiatric inpatients," and nine criterion-level risk domains. Five members of the hospital development team reviewed the round-2 Delphi ratings, qualitative findings, and clinical scoring boundaries. They resolved differences in a consensus meeting and entered one group judgment matrix into Yaahp 12.5. Because the individual working sheets were not retained, we report only the group-consensus matrix and do not attempt to reconstruct or aggregate individual matrices. Consistency was calculated as $CI = (\lambda_{max} - n)/(n - 1)$ and $CR = CI/RI$, where n is the matrix order and RI is the random consistency index. $CR < .10$ indicated acceptable consistency.

The group judgment matrix met the consistency criterion (CR = .024). Each first-level weight was multiplied by 9 to derive a standardized domain coefficient; the nine coefficients summed to 9. Within each domain, the assessor selected one mutually exclusive descriptor scored from 0 to 3. The weighted raw total was the sum of each standardized coefficient multiplied by its domain score, giving a theoretical maximum of 27. First-level weights and standardized coefficients were retained to four decimal places for calculation, and the weighted total was rounded to the nearest integer for clinical reporting. The Youden-optimal statistical cut point was ≥12, while ≥14 was used as the exploratory clinical high-risk cut point. Both the scoring rule and cut points require independent external validation before clinical use.

2.5 Phase 3: Preliminary Clinical Validation

Clinical validation was conducted in two psychiatric wards at Chuxiong Prefecture Second People's Hospital, a tertiary psychiatric specialty hospital in southwest China. Psychiatric inpatients were eligible if they were aged 18 years or older, could cooperate with the assessment, and were expected to remain in hospital for at least 1 month. Exclusion criteria were an acute physical emergency requiring transfer to another department, inability to provide informed consent, or discharge within 48 hours of admission. Ward nurses or research-team members first recorded all assessments and outcomes on paper study forms and later transcribed the data manually into de-identified Excel spreadsheets for analysis.

A total of 156 inpatients were screened. Nineteen were excluded because they did not meet clinical criteria: acute physical illness requiring transfer (n=8), expected length of stay shorter than 1 month (n=6), or severe cognitive impairment preventing assessment (n=5). Of 137 initially eligible patients, seven declined participation or did not sign informed consent. The final sample included 130 consenting patients. Three subsamples were used: an inter-rater agreement subsample of 20 patients nested within the convergent-validity sample, a convergent-validity sample of 80 patients, and an independent prospective sample of 50 newly admitted patients followed for 1 month. The 20 inter-rater agreement cases did not overlap with the 50 prospective cases, and the 80 convergent-validity cases did not overlap with the prospective cases (Figure 1). Paper forms were retained at the study site; only de-identified manually entered spreadsheets were used for the analyses reported here.

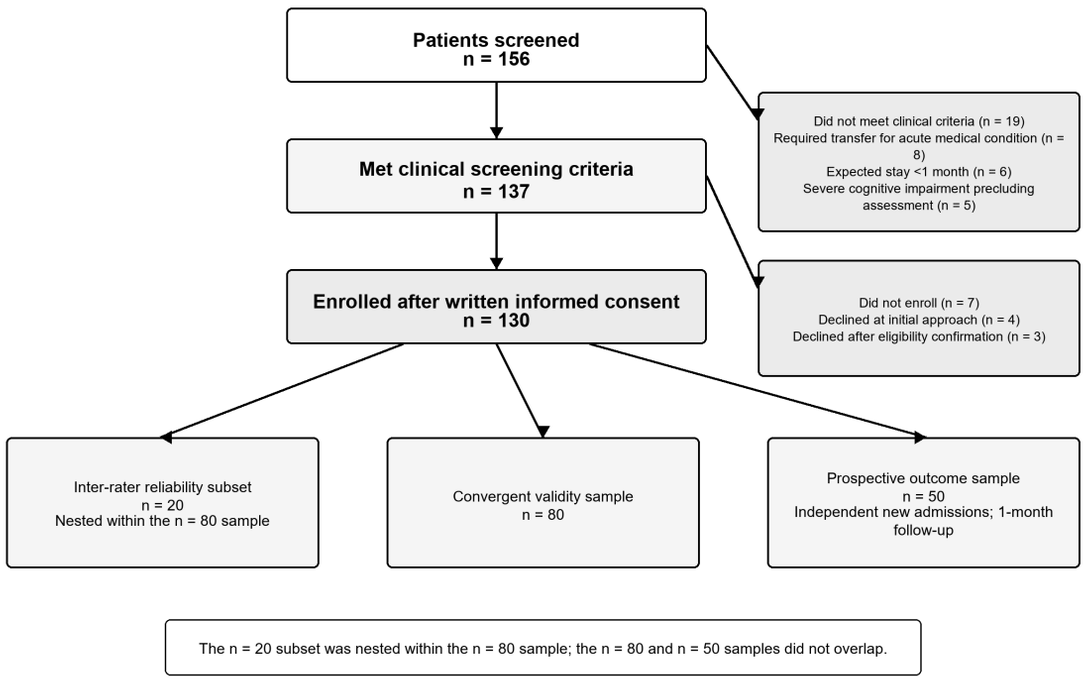

Figure 1. Patient screening and enrollment flowchart.

2.5.1 Inter-Rater Agreement

Two trained psychiatric nurses independently assessed 20 randomly selected patients covering different risk levels on the same day without discussion. Before formal assessment, both nurses received standardized training, including explanation of the scoring descriptors, scoring rules, three standardized patient simulation exercises, and an agreement check. Intraclass correlation coefficient (ICC[2,1]) was calculated using a two-way random-effects model with absolute agreement. Weighted kappa with linear weights was calculated for total-score risk categories based on rounded clinical reporting scores: low risk, 6 or lower; moderate risk, 7 to 13; and high risk, 14 or higher. Exact agreement rate and mean absolute score difference were also reported.

2.5.2 Content Validity

Eight senior psychiatric nursing experts who had not participated in the Delphi consultation independently rated the content relevance of the 36 scoring descriptors using a 4-point scale (1 = not relevant, 4 = highly relevant). Item-level content validity index (I-CVI) and scale-level content validity index average (S-CVI/Ave) were calculated. The Delphi panel included 15 experts in each round, whereas the eight experts in this section were an independent content-validity panel. Polit et al. (2007) recommended a threshold of I-CVI ≥ .78 for eight experts; however, with eight raters, possible I-CVI values are .625, .750, .875, and 1.000. Therefore, this study used I-CVI ≥ .875, equivalent to at least seven of eight experts rating the descriptor as relevant, as the operational threshold.

2.5.3 Convergent Validity

Eighty patients completed both the new tool and the Morse Fall Scale. Pearson correlation was used to examine convergence between total scores. Because both instruments used ordered scoring and scores might not meet normality assumptions, Spearman rank correlation was added as a sensitivity analysis. The Morse Fall Scale was treated as a comparator instrument rather than as a gold standard for true fall outcomes.

2.5.4 Exploratory Prospective Risk Discrimination

Fifty newly admitted patients were assessed within 24 hours of admission and followed for 1 month. A fall was defined as any unplanned descent to the floor or a lower level. The research team monitored outcomes through weekly reviews of nursing records and adverse-event reports, and nurses reported any fall within 24 hours. The outcome was whether a patient experienced at least one fall during follow-up. Outcome assessors were blinded to baseline tool scores. Scores did not alter routine preventive interventions; all patients received standard ward fall-prevention care.

An exploratory ROC curve was used to identify the statistical best cut point by the Youden index. The area under the curve (AUC) 95% confidence interval was estimated using the Hanley-McNeil method (Hanley & McNeil, 1982). For the

clinically selected high-risk cut point, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (LR+), and negative likelihood ratio (LR−) were calculated. Exact binomial confidence intervals were used for sensitivity, specificity, PPV, and NPV.

2.6 Statistical Analysis

Excel 2019, SPSS 26.0, and Yaahp 12.5 were used. Expert positivity was described by valid response rate, expert authority by judgment basis coefficient, familiarity coefficient, and authority coefficient, and expert coordination by CV and Kendall's coefficient of concordance. Inter-rater agreement was described using ICC(2,1), weighted kappa, exact agreement rate, and mean absolute score difference. Convergent validity was assessed using Pearson correlation and Spearman sensitivity analysis. Exploratory risk discrimination was described using ROC analysis and AUC. Fisher's exact test was specified for comparisons between the high-risk and low-to-moderate-risk groups when expected cell counts were lower than 5. Risk ratio (RR) and 95% confidence interval were reported. All tests were two-sided with $\alpha = .05$.

3. Results

3.1 Qualitative Findings

Four psychiatric-specific themes were identified: adverse effects of antipsychotic medication, symptom-related risk, altered consciousness and cognition, and impaired toileting or excretion ability. These themes were translated into four psychiatric-specific risk domains through a mapping process that linked interview evidence with operational scoring boundaries (Table 1).

3.2 Delphi Expert Characteristics, Positivity, Authority, and Coordination

The 15 Delphi experts were aged 32 to 58 years. Most had at least 20 years of work experience (80.0%) and senior professional titles (80.0% associate senior or higher); 13.3% held a master's degree and the remainder held bachelor's degrees. Both Delphi rounds had a 100% valid response rate. The authority coefficient was 0.846 ± 0.115, with judgment basis coefficient of 0.812 ± 0.187 and familiarity coefficient of 0.893 ± 0.099 (Table 2).

Table 2. Delphi Expert Panel Characteristics, Response Rates, and Authority

Characteristic	Category or indicator	Value
Age	31–40 years	3 (20.0%)
Age	41–50 years	7 (46.7%)
Age	>50 years	5 (33.3%)
Sex	Female	12 (80.0%)
Sex	Male	3 (20.0%)
Work experience	10–19 years	3 (20.0%)
Work experience	≥20 years	12 (80.0%)
Professional title	Intermediate	3 (20.0%)
Professional title	Associate senior	9 (60.0%)
Professional title	Senior	3 (20.0%)
Education	Bachelor's degree	13 (86.7%)
Education	Master's degree	2 (13.3%)
Professional area	Clinical nursing/nursing management	8 (53.3%)
Professional area	Psychiatry	4 (26.7%)
Professional area	Clinical pharmacy	2 (13.3%)
Professional area	Rehabilitation medicine	1 (6.7%)
Round 1 response	Valid response rate	15/15 (100.0%)
Round 2 response	Valid response rate	15/15 (100.0%)
Judgment basis coefficient	Ca	0.812 ± 0.187
Familiarity coefficient	Cs	0.893 ± 0.099
Authority coefficient	Cr	0.846 ± 0.115

Note: Kendall's W values were recalculated from the corrected Delphi Excel rating matrices using tie-corrected Friedman tests. The 45-cell analysis included nine domains across five rating dimensions. AHP = analytic hierarchy process; CR = consistency ratio; CV = coefficient of variation. The reported AHP CR refers to the group-consensus matrix.

Expert ratings were high across both Delphi rounds. In the first round, the 45 scoring cells had means of 4.33-5.00, CVs of .00-.23, and full-score frequencies of 60.0%-100.0%; the tie-corrected Friedman analysis for all 45 scoring cells yielded Kendall's W=.106, chi-square(44)=70.00, p=.008. In the corrected second-round rating sheet, means were 4.53-5.00, CVs were .00-.14, and full-score frequencies were 60.0%-100.0%; the corresponding Kendall's W was .106, chi-square(44)=70.06, p=.007. When only the nine first-level domains' importance ratings were analysed, W values were .139 in round 1 (chi-square(8)=16.71, p=.033) and .062 in round 2 (chi-square(8)=7.41, p=.493). Taken together, the Delphi process demonstrated high absolute endorsement of the candidate domains and descriptors, but rank-order concordance was limited, probably because ratings clustered at the upper end of the scale and included many ties (Table 3).

Table 3. Delphi Coordination and AHP Weighting Summary

Indicator	Round 1	Round 2
Kendall's W (45 scoring cells)	.106	.106
chi-square (df=44)	70.00	70.06
p value	.008	.007
Kendall's W (9 domain importance ratings)	.139	.062
chi-square (df=8)	16.71	7.41
p value	.033	.493
CV range for 45 scoring cells	.00-.23	.00-.14
Full-score frequency range	60.0%-100.0%	60.0%-100.0%
Group AHP matrix CR	-	.024

Note: Standardized coefficient = first-level weight x 9. Total score = sum of standardized coefficient x domain score. Maximum theoretical score = 27. Rounded clinical reporting scores were categorized as low risk (6 or lower), moderate risk (7 to 13), and high risk (14 or higher). These cut points have not been externally validated.

3.3 Scoring Structure and Weights

The final tool contained nine risk domains and 36 mutually exclusive ordered scoring descriptors. Second-round importance means for the descriptors ranged from 4.67 to 5.00, and CVs ranged from .00 to .15. The group AHP judgment matrix had CR = .024, indicating acceptable consistency. The AHP weights reflected expert-assigned relative importance and should not be interpreted as observed predictive contributions in outcome data. The full scoring structure is shown in Table 4 and Supplementary File S3.

Table 4. Final Scoring Structure and Weights

Risk domain	Score	Ordered scoring descriptors	First-level weight	Standardized coefficient
Previous falls (past 3 months)	0/1/2/3	No fall / one fall / two falls / three or more falls or fall with serious injury	0.1239	1.1151
Comorbid physical disease diagnosis	0/1/2/3	No physical disease / one low-risk stable chronic disease / one moderate-risk disease or at least two low-risk diseases / at least one high-risk disease or at least two moderate-risk diseases	0.1266	1.1394
Transfer and mobility	0/1/2/3	Moves freely / slow or mildly unstable without assistance / clearly unstable, occasionally needs wall support or assistance / unable to walk or transfer without assistance	0.1200	1.0800
Altered consciousness and cognition	0/1/2/3	Alert and cooperative / mild drowsiness or sedation / confusion or stupor / delirium	0.1189	1.0701
Toileting/excretion ability	0/1/2/3	Independent without urinary frequency or nocturnal toileting / urinary frequency or nocturnal toileting at least once / incontinence or requiring assistance with toileting / urgent incontinence or nocturnal toileting at least three times	0.1189	1.0701
High-risk medication	0/1/2/3	No high-risk medication / one high-risk medication / two high-risk medications / three or more high-risk medications	0.1089	0.9801
Age	0/1/2/3	18–64 years / 65–74 years / 75–84 years / ≥85 years	0.1089	0.9801
Adverse effects of antipsychotic medication	0/1/2/3	No adverse effects / adverse effects not affecting activity / adverse effects clearly affecting activity / adverse effects causing activity limitation or requiring intervention	0.1089	0.9801

Risk domain	Score	Ordered scoring descriptors	First-level weight	Standardized coefficient
Symptom-related risk	0/1/2/3	No symptoms or no symptom-related risk / symptoms causing distractibility but patient responds to safety reminders / symptoms causing predictable risky behavior requiring observation / symptoms causing uncontrollable risky behavior requiring restraint or intensive protection	0.0650	0.5850

Note: Standardized coefficient = first-level weight × 9. Total score = Σ(standardized coefficient × domain score). Maximum theoretical score = 27. Rounded clinical reporting scores were categorized as low risk (≤6), moderate risk (7–13), and high risk (≥14). These cut points have not been externally validated.

3.4 Preliminary Clinical Validation

In the inter-rater subsample of 20 patients, the two raters assigned identical total scores in 65.0% of cases (13/20); the mean absolute difference was 0.60 ± 0.88 points (range, 0–2). ICC(2,1) was .986, indicating high absolute agreement. For the prespecified categories (low, ≤6; moderate, 7–13; high, ≥14), category agreement was 80.0% (16/20), unweighted κ was .697, and linearly weighted κ was .762 (bootstrap 95% CI, .492–.944).

Table 5. Baseline Characteristics of the Convergent-Validity and Prospective Follow-Up Samples

Characteristic	Convergent-validity sample (n = 80)	Prospective follow-up sample (n = 50)
Age, years, mean ± SD	56.3 ± 14.2	55.8 ± 13.6
18–44 years	22 (27.5%)	15 (30.0%)
45–59 years	23 (28.8%)	14 (28.0%)
60–74 years	24 (30.0%)	14 (28.0%)
≥75 years	11 (13.8%)	7 (14.0%)
Female	44 (55.0%)	27 (54.0%)
Schizophrenia	42 (52.5%)	25 (50.0%)
Affective disorder	18 (22.5%)	12 (24.0%)
Organic mental disorder	12 (15.0%)	8 (16.0%)
Other psychiatric diagnosis	8 (10.0%)	5 (10.0%)
Number of psychiatric admissions, mean ± SD	2.8 ± 2.1	2.6 ± 1.9
Fall history in past 3 months	22 (27.5%)	14 (28.0%)
Fall history in past year	32 (40.0%)	21 (42.0%)
Current psychiatric medication	58 (72.5%)	36 (72.0%)
Antipsychotic medication	48 (60.0%)	30 (60.0%)
Sedative-hypnotic medication	28 (35.0%)	17 (34.0%)
Antidepressant medication	18 (22.5%)	11 (22.0%)
Mood stabilizer	10 (12.5%)	6 (12.0%)

Note: Values are n (%) unless otherwise indicated. Diagnoses and medication categories could overlap.

For content validity, I-CVI values ranged from .75 to 1.00. Thirty-two of 36 descriptors (88.9%) met the operational threshold of I-CVI ≥ .875. Four descriptors had I-CVI = .75, distributed across previous falls, altered consciousness and cognition, high-risk medication, and toileting/excretion ability. These descriptors were retained after research-team review because they had Delphi importance means of at least 4.67, were supported by qualitative expert feedback, and were considered necessary for domain completeness. S-CVI/Ave was .92 (Table 6).

Table 6. Content Validity Index Summary by Risk Domain

Risk domain	Number of descriptors	I-CVI range	Descriptors meeting I-CVI ≥ .875	Descriptors below threshold	Mean I-CVI
Previous falls	4	.75–1.00	3	1	.906
Comorbid physical disease diagnosis	4	.875–1.00	4	0	.969
Transfer and mobility	4	.875–1.00	4	0	.969
Altered consciousness and cognition	4	.75–1.00	3	1	.906

Risk domain	Number of descriptors	I-CVI range	Descriptors meeting I-CVI ≥ .875	Descriptors below threshold	Mean I-CVI
Adverse effects of antipsychotic medication	4	.875–.875	4	0	.875
Symptom-related risk	4	.875–1.00	4	0	.969
Age	4	.875–.875	4	0	.875
High-risk medication	4	.75–1.00	3	1	.906
Toileting/excretion ability	4	.75–1.00	3	1	.906
Total (S-CVI/Ave)	36	.75–1.00	32	4	.92

Note: With eight experts, possible I-CVI values are .625, .750, .875, and 1.000. The operational threshold was .875, equivalent to at least seven of eight experts rating the descriptor as relevant.

In the convergent-validity sample, scores on the new tool correlated strongly with Morse Fall Scale scores (Pearson $r = .848$, $p < .001$; Spearman $\rho = .777$, $p < .001$). Risk-category agreement between the instruments was 78.8%, with an unweighted κ of .647. Because the same responsible nurse completed both instruments during the same assessment period, common-method bias is possible; the Morse Fall Scale was therefore treated as a comparator rather than a gold standard.

All 50 patients completed 1-month follow-up, with no loss to follow-up. In the prospective exploratory sample, ROC AUC was .876, with a Hanley-McNeil 95% CI of .673 to 1.000. The confidence interval was wide because only five fall events occurred. The Youden-optimal cut point was a clinical reporting score of 12 or higher, with sensitivity 1.000 and specificity .756. Combining clinical interpretability with the prespecified risk categories, a score of 14 or higher was used as the exploratory high-risk cut point. At this cut point, sensitivity was .600, specificity was .800, PPV was .250, NPV was .947, LR+ was 3.00, and LR- was .50. Among 50 followed patients, falls occurred in 3 of 12 high-risk patients (25.0%) and 2 of 38 low-to-moderate-risk patients (5.3%). Fisher's exact test was not statistically significant (two-sided p=.082); the risk ratio was 4.75 (95% CI .90 to 25.17). The Cochran-Armitage trend test across low, moderate, and high categories gave p=.036. These findings support only an initial risk-discrimination signal (Figure 2).

Table 7. Preliminary Reliability, Convergent Validity, and Risk-Discrimination Findings

Analysis	Indicator	Result
Inter-rater agreement (n=20)	Exact agreement	65.0% (13/20)
Inter-rater agreement (n=20)	Mean absolute score difference	0.60 +/- 0.88 points (range 0-2)
Inter-rater agreement (n=20)	ICC(2,1)	.986
Inter-rater agreement (n=20)	Category agreement	80.0% (16/20)
Inter-rater agreement (n=20)	Unweighted kappa	.697
Inter-rater agreement (n=20)	Linearly weighted kappa	.762 (bootstrap 95% CI .492-.944)
Convergent validity (n=80)	Pearson correlation with Morse Fall Scale	r=.848, p<.001
Convergent validity (n=80)	Spearman sensitivity analysis	rho=.777, p<.001
Convergent validity (n=80)	Risk-category agreement / kappa	78.8% / .647
Prospective exploratory sample (n=50)	Falls in high-risk group (>=14)	3/12 (25.0%)
Prospective exploratory sample (n=50)	Falls in low-to-moderate-risk group	2/38 (5.3%)
Prospective exploratory sample (n=50)	Fisher exact test	p=.082
Prospective exploratory sample (n=50)	Risk ratio	4.75 (95% CI .90-25.17)
ROC analysis (n=50)	AUC	.876 (95% CI .673-1.000)
ROC analysis (n=50)	Youden-optimal cut point	>=12; sensitivity 1.000; specificity .756
Clinical high-risk cut point (>=14)	Sensitivity / specificity	.600 / .800
Clinical high-risk cut point (>=14)	PPV / NPV	.250 / .947

Note: AUC confidence interval was estimated using the Hanley-McNeil method. Exact binomial confidence intervals were used for sensitivity, specificity, PPV, and NPV. The linearly weighted kappa confidence interval was obtained by bootstrap resampling of the 20 paired assessments. Likelihood-ratio confidence intervals were not calculated because of the small number of events.

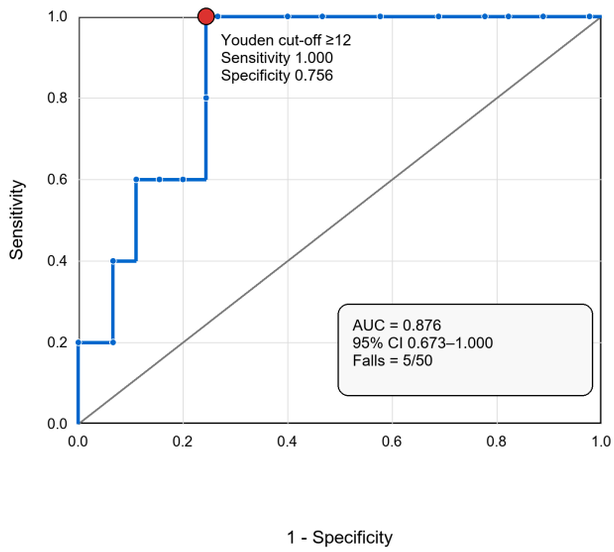

Figure 2. Receiver operating characteristic curve for the exploratory prospective sample.

4. Discussion

4.1 Principal Findings

This exploratory sequential mixed-methods study developed a psychiatric inpatient fall risk assessment tool containing nine risk domains and 36 mutually exclusive ordered scoring descriptors. The tool incorporated conventional fall-risk factors and psychiatric-specific domains derived from qualitative interviews. The group AHP matrix had acceptable consistency, content validity was acceptable, total-score inter-rater agreement was high, convergence with the Morse Fall Scale was strong, and the short-term risk-discrimination signal was plausible but unstable. Delphi ratings supported high absolute endorsement of the domains and descriptors, although rank-order concordance was weak and should be interpreted in light of ceiling effects and tied scores.

The study's most important contribution is the integration of psychiatric ward experience into operational scoring descriptors. In psychiatric units, fall risk is not limited to physical frailty or mobility impairment. Sedation, orthostatic symptoms, extrapyramidal reactions, hallucination-driven behavior, delusional fear, impaired insight, fluctuating consciousness, and urgent toileting can alter risk within short time windows. By translating interview evidence into scoring descriptors, the tool may better match the clinical reasoning used by psychiatric nurses than instruments designed primarily for general hospital settings. This emphasis is consistent with qualitative evidence that mental health nurses combine structured information with contextual clinical judgment when assessing patient risk (Conlon et al., 2024).

4.2 Interpretation of Delphi Consensus and AHP Weighting

High mean ratings, low coefficients of variation, high full-score frequencies, and fewer revision comments in round 2 indicated strong absolute endorsement of the domains and scoring descriptors. Rank-order concordance was nevertheless modest, particularly for the nine first-level domains. The clustering of ratings at the upper end of the scale and the resulting tied ranks offer a plausible explanation for this pattern. The Delphi findings therefore support content development and refinement rather than a strong hierarchy among domains. AHP made the development team's group judgments explicit, but future outcome-based validation should test the empirical contribution of each domain.

The tool's formative structure also determines which measurement properties are most informative. Unlike reflective psychometric scales, formative clinical risk tools combine heterogeneous indicators to define a risk profile. When domains are conceptually distinct, internal consistency coefficients such as Cronbach's alpha are not the primary evidence and may be misleading. Evaluation should instead emphasize content coverage, scoring clarity, inter-rater agreement, calibration, discrimination, and clinical utility.

4.3 Content Validity and Scoring Descriptors

The S-CVI/Ave of .92 indicated acceptable overall content validity. Four descriptors fell below the operational I-CVI

threshold of .875, although each was rated as relevant by six of eight content-validity experts. The research team retained them because deletion would have reduced coverage of their respective risk domains. These descriptors nevertheless require clearer operational definitions and reassessment. Examples, anchor descriptions, or brief clinical vignettes may improve scoring consistency and perceived relevance.

4.4 Preliminary Risk Discrimination

The exploratory AUC of .876 appears promising but should be interpreted cautiously. With only five fall events, the 95% confidence interval was wide, and the estimate is vulnerable to instability. The Youden-optimal cut point of 12 or higher achieved sensitivity 1.000 and specificity .756 in this sample, but it was selected after inspecting the same data and was not externally validated. The clinical high-risk cut point of 14 or higher yielded specificity .800 but sensitivity only .600. LR+ of 3.00 provides limited to moderate support for identifying higher-risk patients, whereas LR- of .50 is insufficient to rule out fall risk. PPV was low because the fall-event rate was low, and NPV should not be interpreted as proof of safety. The nonsignificant Fisher's exact test, trend result based on few events, and wide risk-ratio confidence interval reinforce the need for larger external validation.

4.5 Clinical Implications

At this stage, the tool is a preliminary clinical assessment framework rather than a validated decision rule. It may help psychiatric nurses consider psychiatric symptoms, medication adverse effects, cognition, and toileting behavior alongside conventional fall-risk factors. Classification as low or moderate risk should not be used to reduce routine fall-prevention precautions. Any use before external validation should include staff training, documentation of the scoring rationale, and ongoing outcome monitoring. Future implementation work should also involve patients, whose experiences suggest that risk assessment can influence participation, perceived responsibility, and therapeutic relationships (Nyman et al., 2022).

4.6 Limitations

Several limitations should be noted. First, all data came from two wards in a single tertiary psychiatric hospital in southwest China, limiting external generalizability. Second, the clinical validation samples were small, and the prospective analysis included only 50 patients and five fall events. Third, although ICC and weighted kappa were favorable, the inter-rater agreement sample was small. Fourth, four scoring descriptors did not meet the operational I-CVI threshold and require revision and reassessment. Fifth, the statistical cut point of 12 or higher and the clinical high-risk cut point of 14 or higher were not locked in an independent validation sample. Sixth, AHP weights were derived from a group consensus matrix; future studies may preserve individual judgment records to further improve traceability of the weighting process. Seventh, data were collected on paper forms and then manually transcribed into de-identified spreadsheets; future studies could use double-entry verification or electronic data capture to further reduce transcription error. Finally, no independent external validation sample was included, and all validation results should be considered internal and exploratory.

4.7 Future Research

Future work should use multicentre prospective cohorts with enough fall events, ideally 50–100, to estimate discrimination, calibration, and confidence intervals more reliably. The scoring rule and risk cut points should be prespecified. Direct comparison with existing psychiatric fall-risk instruments is also needed, together with evaluation of incremental prediction, calibration, decision-curve performance, usability, time burden, and effects on fall-prevention workflows.

5. Conclusion

Qualitative interviews, Delphi consultation, and a group-consensus AHP matrix were used to develop a fall risk assessment tool for psychiatric inpatients. The final tool comprised nine risk domains and 36 mutually exclusive ordered scoring descriptors. Preliminary findings supported content relevance, high inter-rater agreement for total scores, and convergence with the Morse Fall Scale. Short-term risk discrimination was promising but unstable, and Delphi rank-order concordance was limited by ceiling effects. The tool requires larger multicentre prospective external validation before it can be recommended for clinical decision support.

Acknowledgments

The authors thank the participating patients, nurses, and Delphi experts for their time and clinical insights.

Author Contributions

Xiuqiong Gao contributed to study conception and design, the literature search, qualitative interviews, data collection and

statistical analysis, supervision, critical review, preparation of the first manuscript draft and manuscript revision. Qiongxian Zhao, Huifang Li, and Chengbao Zhu contributed to study design, quality control, expert recruitment, and funding support. Feng Guo and Wenhui Xi contributed to data collection and statistical analysis. All authors reviewed and approved the final manuscript.

Funding

This work was supported by the Scientific Research Fund of Chuxiong Medical College (Project No. 2025YYXM17). The funder had no role in the study design, data collection, analysis, interpretation, or preparation of the manuscript.

Ethics Approval and Consent to Participate

The Ethics Committee of Chuxiong Prefecture Second People's Hospital approved the study on April 18, 2025 (Approval No. 2025-07-Ke). All participating patients and nurses provided written informed consent.

Data Availability Statement

The de-identified summary data used for the analyses are available from the corresponding author on reasonable request, subject to institutional ethics and privacy restrictions. Original paper assessment forms are retained by the participating hospital and are not publicly posted.

Declaration of Interest

The authors declare no conflicts of interest.

References

[1] Aczel, J., & Saaty, T. L. (1983). Procedures for synthesizing collective judgements. Journal of Mathematical Psychology, 27(2), 165–175.

[2] Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. Psychological Bulletin, 110(2), 305–314.

[3] Chen, H., Li, X., & Wang, Y. (2023). Construction of a nursing risk assessment index system for psychiatric inpatients. Chinese Nursing Management, 23(1), 63–68. [in Chinese]

[4] Conlon, D., Raeburn, T., & Wand, T. (2024). Mental health risk assessments of patients, by nurses working in mental health settings: A qualitative study using cognitive continuum theory. Issues in Mental Health Nursing, 45(5), 488–497. https://doi.org/10.1080/01612840.2024.2320731

[5] Duan, G., Xie, S., Zhang, Y., et al. (2026). Status and differences in fall-risk perception among psychiatric inpatients. Psychology Monthly, 21(1), 33–36, 41. [in Chinese]

[6] Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic curve. Radiology, 143(1), 29–36.

[7] Huang, H., Su, F., Huang, Q., et al. (2023). Interpretation of key points from the World Guidelines for Falls Prevention and Management for Older Adults. Military Nursing, 40(5), 113–116. [in Chinese]

[8] Janse van Rensburg, E., & van der Wath, A. (2020). Risk assessment in mental health practice: An integrative review. Issues in Mental Health Nursing, 41(11), 995–1003. https://doi.org/10.1080/01612840.2020.1756011

[9] Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155–163.

[10] Li, Y., & Chen, W. (2018). Application status of the Delphi method in nursing management. Nursing Research, 32(21), 3339–3341. [in Chinese]

[11] Liu, S., Jiang, H., Li, K., et al. (2020). Construction of quality indicators for hematology specialty nursing. Chinese Journal of Nursing, 55(4), 574–578. [in Chinese]

[12] Nyman, M., Hofvander, B., Nilsson, T., & Wijk, H. (2022). "You should just keep your mouth shut and do as we say": Forensic psychiatric inpatients' experiences of risk assessments. Issues in Mental Health Nursing, 43(2), 137–145. https://doi.org/10.1080/01612840.2021.1956658

[13] Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Research in Nursing & Health, 30(4), 459–467.

[14] Rao, W. W., Zong, Q. Q., Lok, G. K. I., et al. (2018). Prevalence of falls in adult and older adult psychiatric patients in China: A systematic review and comprehensive meta-analysis of observational studies. Psychiatry Research, 266, 18–25.

[15] Shen, Y. H., Hsieh, C. C., Lee, M. T., et al. (2021). Simplifying and testing the psychometric Psychiatric Patients' Fall Risk Scale: An analysis of one-year admissions. Healthcare, 9(9), 1119.
[16] Shi, Y., & Zhu, Q. (2015). Application of the Morse Fall Scale in psychiatric settings. Today Nurse, 2015(8), 116–118. [in Chinese]
[17] Siafis, S., Wu, H., Wang, D., et al. (2023). Antipsychotic dose, dopamine D2 receptor occupancy and extrapyramidal side-effects: A systematic review and dose-response meta-analysis. Molecular Psychiatry, 28(8), 3267–3277.
[18] World Health Organization. (2021, April 26). Falls. https://www.who.int/news-room/fact-sheets/detail/falls
[19] Zhang, H., Jing, L., Xu, Y., et al. (2023). Weight analysis of a comprehensive evaluation index system for hospice care services based on analytic hierarchy process. Chinese Journal of Health Policy, 16(1), 52–59. [in Chinese]
[20] Zhang, M., Xu, D., Gao, J., et al. (2022). Construction of an assessment form for protective restraint implementation in psychiatric departments based on the Delphi method. Nursing Research, 36(14), 2466–2470. [in Chinese]