Received: 19 August 2022 | Revised: 17 January 2024 | Accepted: 24 January 2024

DOI: 10.1111/eip.13516

ORIGINAL ARTICLE

WILEY

The Italian version of the 16-item Prodromal Questionnaire (PQ-16) and its psychometric features in help-seeking ultra-high-risk subjects and in the general population

Alberto Parabiaghi[1] | Alessandro Alberto Rossi[2,3] | Anna Castelnovo[4,5,6] | Lorenzo del Fabro[7,8] | Stefania Mannarini[2,3] | Mauro Emilio Percudani[8,9] | for the CCM2013 Project team (corporate authorship)

[1]Department of Health Policy, Unit for Quality of Care and Rights Promotion in Mental Health, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan, Italy

[2]Department of Philosophy, Sociology, Education, and Applied Psychology, Section of Applied Psychology, University of Padova, Padova, Italy

[3]Interdepartmental Center for Family Research, University of Padova, Padova, Italy

[4]Sleep Medicine Unit, Neurocenter of Southern Switzerland, Lugano, Switzerland

[5]Faculty of Biomedical Sciences, Università della Svizzera Italiana, Lugano, Switzerland

[6]University Hospital of Psychiatry and Psychotherapy, University of Bern, Bern, Switzerland

[7]Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy

[8]Department of Neurosciences and Mental Health, IRCCS Fondazione Ca' Granda Ospedale Maggiore Policlinico, University of Milan, Milan, Italy

[9]Department of Mental Health and Addiction Services, ASST Grande Ospedale Metropolitano "Niguarda", Milan, Italy

Correspondence
Alberto Parabiaghi, Department of Health Policy, Unit for Quality of Care and Rights Promotion in Mental Health, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan, Italy.
Email: alberto.parabiaghi@marionegri.it

Abstract

Background: Increasing attention to the early stages of psychosis and the identification of symptomatic prodromal states have led to the development of a growing number of screening tools. The 16-item version of the Prodromal Questionnaire (PQ-16) is a worldwide used self-administered tool for this purpose. However, to date, fundamental psychometric properties of PQ-16 were not thoroughly investigated. This study aimed to examine the structural validity, measurement invariance, reliability and other psychometrical properties of the Italian version of the PQ-16 (iPQ-16) in help-seeking individuals and in the general population.

Methods: The iPQ-16 was administered to 449 young outpatients attending six community mental health services and to 318 control participants enrolled in educational environment. Confirmatory factor analyses (CFAs), measurement invariance (MI) between the help-seeking group and the general population sample, convergent validity, test–retest reliability, internal consistency, and prevalence analyses were performed. Lastly, the validity of the adopted PQ-16 cut-offs through Receiver Operating Characteristic (ROC) curves plotted against CAARMS diagnoses was also tested.

Team involved in CCM2013 Project listed in 'Acknowledgements' section.

Results: CFAs confirmed the single-factor structure for the iPQ-16 and scalar MI was reached. The iPQ-16 showed high internal consistency, test–retest reliability, convergent validity, and acceptable diagnostic accuracy. ROC analysis suggested a score of ≥4 as best cut-off.

Conclusions: The iPQ-16 represents a valid and reliable questionnaire for the assessment of high mental risk in both Italian outpatients and general student population. It has good psychometric properties and is easy to implement as UHR screening for clinical as well as research purposes.

KEYWORDS
PQ-16, prodromal symptoms, psychosis, scale validation, ultra high risk

1 | INTRODUCTION

The constructs of the 'Ultra-High Risk' (UHR) state and the 'At-Risk Mental State' (ARMS) are garnering attention for their role in identifying individuals who may be at high risk for psychosis (Fusar-Poli et al., 2013; Yung et al., 2004). These terms are associated with individuals who may benefit from early intervention services, even in the absence of a definitive psychosis diagnosis. Indeed, the process of early diagnosis and ensuring timely access to appropriate services remains a challenge. Established interviews such as the Structured Interview for Prodromal States (SIPS) (Miller et al., 2003) and the Comprehensive Assessment of At-Risk Mental States (CAARMS) (Yung et al., 2005) are reliable tools for diagnosing the UHR state. However, their administration requires extensive training and is time-consuming, which can hinder early identification of cases in the general population beyond specialized early intervention clinics (Kline & Schiffman, 2014). To enhance the efficiency of UHR identification, a two-stage model incorporating self-report screening followed by a clinical interview has been proposed as an accurate and efficient approach. Consequently, in recent years, several screening instruments have been developed and validated to assess the risk of developing psychosis in the help-seeking population (Kline et al., 2012) (see Supplementary S1).

The PQ-16 is derived from the Prodromal Questionnaire (PQ-92), which is a 92-item self-report measure designed to identify UHR individuals in mental health services using a two-stage screening process in conjunction with a subsequent clinical assessment (Kotzalidis et al., 2017; Loewy et al., 2005). The PQ has shown utility as a preliminary screening tool for UHR in various settings, as demonstrated by several studies (Kline & Schiffman, 2014). To enhance efficiency and accuracy, a shorter version of the PQ-92, known as the PQ-B, was developed. The PQ-B consists of 21 items and includes a Likert scale to assess the frequency of each symptom and its associated distress or impairment (Kotzalidis et al., 2017; Loewy et al., 2005). The PQ-16, a 16-item version of the Prodromal Questionnaire, was specifically designed to identify individuals with attenuated psychosis syndrome in large help-seeking populations within secondary mental health care services for non-psychotic disorders (Ising et al., 2012). It comprises the most predictive items from the PQ-92: nine hallucination-like items, five delusion-like items, and two negative symptom items, demonstrating high sensitivity (87%) and specificity (87%) in predicting the UHR state.

Recently, the PQ-16 has been translated into Italian (iPQ-16) (Pelizza et al., 2018; Pelizza et al., 2019), which has contributed to promoting its usage in young help-seeking populations. While the iPQ-16 has shown promising and satisfactory psychometric properties, its validation was based on a relatively small sample from a single centre and did not include a factor analysis (either exploratory or confirmatory), assuming 'de facto' the PQ-16's factorial structure. The validation primarily focused on diagnostic accuracy, content validity, convergent validity, and concurrent validity. Only one study has assessed the factorial structure of the PQ-16 (Howie et al., 2022), in an English-speaking sample. However, despite its large sample size and robust statistical analysis, this study was conducted solely on individuals recruited online from the general population and yielded different results from the initial exploratory factor analysis and subsequent confirmatory analysis.

Furthermore, to the best of our knowledge, none of the previous studies assessing the psychometric properties of the PQ-16 have compared the factorial structure between UHR patients and healthy control subjects. They have simply assumed that the model structure is the same for both groups. Additionally, the optimal cut-off for the iPQ-16 to improve concurrent validity and clinical usefulness has yet to be determined through multicenter testing.

Therefore, the aim of the current investigation was to comprehensively examine various psychometric properties of the Italian translation of the PQ-16 to address the gaps in the existing scientific literature. This included assessing its structural validity, internal consistency, test–retest reliability, and, particularly, evaluating its factorial structure at the item level and examining measurement invariance between UHR patients and young adults from the general population.

2 | MATERIALS AND METHODS

2.1 | Participants

This study is part of a National Project funded by the Italian Centre for Disease Prevention and Control aimed at implementing the UHR

paradigm in six departments of mental health in Italy (sited in Lombardy, Liguria, and Tuscany) and developing an integrated approach to address clinical practice and service organization for youth in a broader preventive perspective (the CCM2013 Project) (Parabiaghi et al., 2019).

The overall sample comprised 767 participants. 449 UHR patients [217 males (48.3%) and 232 females (51.7%) with age ranging from 14 to 24 years (*mean* = 20.40, SD = 2.62)] were enrolled during their first psychiatric visit, in different mental health departments: (I) ASST Rhodense, Garbagnate Milanese ($n = 109$; 24.3%), (II) ASST di Lecco, Lecco ($n = 5$; 1.1%), (III) ASST Fatebenefratelli-Sacco, Milano ($n = 11$; 2.4%), (IV) ASST Grande Ospedale Metropolitano 'Niguarda', Milano ($n = 72$; 16.0%), (V) ASL 3 Genovese, Genova ($n = 105$; 23.4%), (VI) Azienda USL 9 Grosseto ($n = 50$; 11.1%), and (VII) ASST Santi Paolo e Carlo ($n = 97$; 21.6%).

In addition, 318 control participants [155 males (48.7%) and 163 females (51.3%) aged from 15 to 24 years (mean = 18.92, SD = 3.02)] taken from a general student population were enrolled in the current study. Specifically, they have been selected from a high school 'Liceo Elio Vittorini' in Milan ($n = 159$; 50%) and from the University of Milan ($n = 159$; 50%).

All the participants were Italian native speakers; aged between 15 and 24. Exclusion criteria included: (A) illiteracy, (B) a diagnosis of psychosis, (C) inability to complete the assessment due to vision and/or (D) cognitive impairments.

Written informed consent was obtained from all the patients and/or their legal caregiver(s). Local ethics committees' approval was not required as per the Italian legislation. The project was approved and funded by the National Centre for Disease Prevention and Control and supported by the Lombardy Region. The CCM2013 Project (supplementary material S2) was an implementation project aimed at transferring best clinical practices into everyday mental health care. Thus, patients data were collected for specific clinical purposes. Routine written informed consent was obtained from each recruited subject, permission for further data use and analysis was implied.

2.2 | Measures

2.2.1 | The Prodromal Questionnaire 16 (PQ-16)

The PQ-16 (Ising et al., 2012) is a self-report screening questionnaire specifically developed to identify individuals at UHR for psychosis (Azzali et al., 2018; de Jong et al., 2018; Ising et al., 2012). It assesses several symptomatic areas, including (A) hallucinations and perceptual aberrations, (B) negative symptoms, and (C) delusions and unusual thought content (Chen, Wang, Heeramun-Aubeeluck, et al., 2014; Savill et al., 2018). Each of the 16 items investigates a specific symptom of psychosis with two parallel response scales (Azzali et al., 2018; de Jong et al., 2018; Ising et al., 2012).The first scale, the 'absence/presence scale' (A/P-scale), assesses the presence of the aforementioned symptoms on a binary response scale (true/false). The second scale, the 'intensity scale' (I-scale), measures the degree of distress provided by each symptom on a 4-point Likert scale (ranging from 0 = no distress to 3 = severe distress). The PQ-16 provides one total score for the A/P scale (ranging from 0 to 16), calculated as the simple sum of each binary item, and one total score for the I-scale (ranging from 0 to 48), calculated as the sum of all individual scores. When using the total symptom scores of the A/P scale, a cut-off threshold of ≥6 was deemed appropriate for general mental health settings (Ising et al., 2012). Meanwhile, using the distress score, a threshold of ≥8 was supported in help-seeking populations within general mental health settings (Chen et al., 2016; Chen, Wang, & Zhao, 2014). A total score of 6 on the A/P scale was considered an appropriate cut-off for distinguishing between healthy individuals and UHR patients (Ising et al., 2012). In adherence to standard guidelines (Beaton et al., 2000; Guillemin et al., 1993), the scale underwent blind back-to-back translation to ensure linguistic and cultural validity. Additionally, the final version of the iPQ-16 was preliminarily tested on a random sample of 10 UHR patients to assess item comprehensibility for the target population, and no further adjustments were required.

2.2.2 | The comprehensive assessment of At-Risk mental states (CAARMS)

The CAARMS (Fusar-Poli et al., 2012; Raballo et al., 2013; Yung et al., 2005) is a semi-structured clinical interview designed for assessing UHR status for psychosis by investigating signs and symptoms of attenuated psychopathology across various domains: (A) 'Positive Symptoms', (B) 'Cognitive Change, Attention and Concentration', (C) 'Emotional Disturbance', (D) 'Negative Symptoms', (E) 'Behavioural Change', (F) 'Motor/Physical Changes', and (G) 'General Psychopathology'. For the present research, in line with previous studies, only the 'Positive Symptoms' subscale was used to determine UHR criteria (Yung et al., 2005). CAARMS categorizes patients into five groups: (I) no vulnerability, (II) vulnerability, (III) attenuated psychosis, (IV) Brief Limited Intermittent Psychotic Symptoms (BLIPS), (V) psychosis. In adherence to established guidelines (Raballo et al., 2013; Yung et al., 2005), the CAARMS assessments were exclusively conducted by specialized clinical psychiatrists who underwent collective supervision sessions during the course of this study.

2.2.3 | The social and occupational functioning assessment scale (SOFAS)

The SOFAS (Goldman et al., 1992) is a hetero-administered scale designed to assess social and occupational functioning according to DSM criteria. The SOFAS score ranges from 0 (indicating poor functioning) to 100 (representing excellent functioning), with a score below 50 serving as an indicator of 'low functioning' (van der Gaag et al., 2012).

2.2.4 | The health of the nation outcome scales measures (HoNOS and HoNOSCA)

The HoNOS measures, including HoNOS (Lora et al., 2001; Wing et al., 1998, 1999) and HoNOSCA (D'Avanzo et al., 2018; Gowers

et al., 1999, 2000), are scales developed for the routine assessment of outcomes in adults and children/adolescents with mental illness, providing comparable results (Pirkis et al., 2005). These scales consist of 12 items on a 5-point Likert scale, aimed at assessing and quantifying patients' progress in mental health, social, and behavioural functioning. A recent literature review demonstrated the sound psychometric properties of both HoNOS and HoNOSCA (Pirkis et al., 2005). The HoNOS assessment was conducted exclusively by specialized clinical psychiatrists who participated in collective supervision sessions.

2.2.5 | The global assessment of functioning (GAF)

The GAF is a scale developed to measure overall psychological disturbance (Jones et al., 1995) and the severity of illness in psychiatry (Aas, 2011). The GAF aims to assess psychological, social, and occupational functioning (Aas, 2011; Startup et al., 2002). It consists of nine items describing behaviours ranging from the absence of psychopathological symptoms to the presence of danger and/or serious suicidal acts with a clear expectation of death (Jones et al., 1995). Each item has a 9-point Likert scale – the lower the score, the more serious the individual functioning (Hall, 1995; Jones et al., 1995). Each GAF assessment was exclusively conducted by specialized clinical psychiatrists who underwent collective supervision sessions.

2.2.6 | The general health Questionnaire-12 (GHQ-12)

The GHQ (Goldberg, 1972; Piccinelli & Politi, 1993) is a self-report questionnaire aimed at assessing mental health. The GHQ consists of 12 items on a 4-point Likert scale. It has a total score (derived from the simple sum of each item), a 'positive items' score, and a 'negative' items score (Piccinelli & Politi, 1993). Due to its brevity and ease of completion, the GHQ has been extensively used in clinical settings and research practice as a screening tool in the general population (Montazeri et al., 2003).

2.3 | Statistical analyses

We utilized R software (R Core Team, 2014; R Core Team, 2017) to conduct statistical analyses, employing several packages, including psych (Revelle, 2018), lavaan (Rosseel, 2012; Rosseel et al., 2015), lme4 (Bates et al., 2015), ordinal (Christensen, 2019), and pROC (Robin et al., 2011).

2.3.1 | Preliminary analysis: Effect of data clustering

We preliminarily assessed the potential effect of hierarchical/multilevel data clustering (1st level: subjects; 2nd level: mental health departments and schools/universities) using methods described in Heck & Thomas, 2015; Hedges et al., 2012; Hox et al., 2018; Peugh, 2010; Sommet & Morselli, 2017. For each item of the A/P-scale and the I-scale, we computed the intraclass correlation coefficient (ICC), adapted for dichotomous and ordinal data using the latent variable approach (Goldstein et al., 2002; Snijders & Bosker, 1999) as described in Anderson et al., 2014; Austin & Merlo, 2017; Merlo et al., 2006. We also calculated the Median Odds Ratio (MOR) (Merlo et al., 2006; Sanagou et al., 2012) and the Median Rate Ratio (MRR) (Austin et al., 2018) for dichotomous and ordinal data. We used the following cutoff criteria as evidence of a clustering effect: ICC > 0.050, MOR > 2, and MRR > 2 (as described in Aguinis et al., 2013; Brown, 2015; Dyer et al., 2005; Geldhof et al., 2014; Grimm et al., 2017; Hayes, 2006; Heck, 2001; O'Connell, 2010; Sommet & Morselli, 2017; Thomas et al., 2005).

2.3.2 | Structural validity

Two parallel structural models were tested via Confirmatory Factor Analysis (CFA): one for the 'A/P-scale' and another for the 'I-scale.' Due to the binary response scaling of the 'A/P-scale' and the ordinal scaling of the 'I-scale', the diagonally weighted least square (DWLS) estimator was employed (Brown, 2015; Hoyle, 2012; Kline, 2016).

Model fit was evaluated by fit indices, with ideal fit criteria (van de Schoot et al., 2012; Yu, 2002) as follows: Satorra-Bentler Chi-square statistics (S-Bχ^2; $p > 0.05$, nonsignificant) (Muthén & Muthén, 1998-2017), Root-Mean-Square Error of Approximation (RMSEA; <0.08) (Barrett, 2007; Browne & Cudeck, 1989; Hu & Bentler, 1999; Steiger, 1990; van de Schoot et al., 2012), Comparative Fit Index (CFI; >0.95) (Bentler, 1990; Brown, 2015; Browne & Cudeck, 1989, 1992; Hu & Bentler, 1999; van de Schoot et al., 2012), and the ratio of χ^2 to the degrees of freedom (χ^2/df < 3).

2.3.3 | Measurement invariance

Measurement invariance (MI) analyses were conducted to assess whether the Italian version of the PQ-16 (both the 'A/P-scale' and the 'I-scale') exhibited invariance between the sample of UHR patients and individuals from the general population (Vandenberg & Lance, 2000).

A single-factor first-order model was fit to the 'A/P-scale' data using the following steps: first, the structural model (Configural Invariance) was established, and then the factor loadings and item thresholds (Metric + Scalar Invariance) were consecutively constrained to be equal between groups (Brown, 2015; Hirschfeld & Von Brachel, 2014; Manzoni et al., 2020; Millsap & Yun-Tein, 2004; Muthén & Asparouhov, 2002; Muthén & Muthén, 1998-2017).

A single-factor first-order model was fit to the 'I-scale' data using the following steps: the structural model (Configural Invariance), factor loadings (λ) (Metric Invariance), and item thresholds (Scalar Invariance) were successively constrained to be equal between groups (Meredith, 1993; Vandenberg & Lance, 2000).

MI was assessed by testing differences in fit indices, with the following criteria for model equivalence: DIFFTEST (equivalent to S-B$\Delta\chi^2$; p-value > 0.050), ΔCFI (<0.010), and ΔRMSEA (<0.015) (Brown, 2015; Cheung & Rensvold, 2002; Muthén & Muthén, 1998-2017). Evidence of model non-invariance was considered if two out of these three indices exceeded the cutoffs and indicated a worse fit.

2.3.4 | Psychometric properties

Items were also tested for their ability to discriminate between subjects with the absence or presence of prodromal symptoms and those with low or high symptom distress (Chiorri, 2011; Ebel, 1965), known as item discriminant power (IDP). For each item of the 'A/P-scale', we computed the z-statistic, Ebel's item discriminative ability (IDA), and its effect size (h) following the methods described by Chiorri (2011) and Ebel (1965). The interpretation of IDA and h effect size was based on Ebel's benchmarks (Chiorri, 2011; Ebel, 1965). For each item of the 'I-scale', we conducted an independent sample t-test and computed its effect size (Cohen's d) (Cohen, 1960, 1988), along with the adjusted item-total correlation ($r_{it\text{-}tot}$), using the procedures outlined in Chiorri (2011), Ebel (1965), Howell (2013), and Tabachnick and Fidell (2014). Cohen's d and $r_{it\text{-}tot}$ were interpreted according to Cohen's benchmarks (Cohen, 1960, 1988).

McDonald's omega (ω) was used as a measure of internal consistency for the 'A/P-scale' and the 'I-scale' for binary and categorical items, respectively (Green & Yang, 2009; McDonald, 1999; Reise et al., 2013).

Test-retest reliability was estimated on a subsample of 40 UHR patients using the two-way mixed intraclass correlation coefficient ($ICC_{consistency}$) for both the 'A/P scale' and the 'I-scale.' This statistic was also used to assess the stability of the PQ-16 diagnosis (Berchtold, 2016; de Vet et al., 2006; Koo & Li, 2016).

Convergent validity was assessed using the Pearson correlation coefficient (Tabachnick & Fidell, 2014), with interpretation based on Cohen's benchmarks (Cohen, 1960).

Additionally, to investigate differences in the PQ-16 scale ('A/P-scale' and 'I-scale') between CAARMS diagnoses (non-UHR patients vs. UHR patients), Mann-Whitney tests and independent sample t-tests were performed, respectively. The strength of these differences was evaluated using Hedges' g (Hedges & Olkin, 1985) and its associated benchmarks (Cohen, 1988).

2.3.5 | Accuracy of the PQ-16 as a screening/diagnostic tool

Lastly, a Receiver Operating Characteristic (ROC) curve was used to assess the accuracy of the iPQ-16 in discriminating between non-UHR patients and UHR patients (Ising et al., 2012; Pepe, 2003; Savill et al., 2018; Zhou et al., 2002). The CAARMS diagnosis (non-UHR patients vs. UHR patients) was used as the external criterion variable, and the iPQ-16 total scores of the 'A/P-scale' and the 'I-scale' were used as the dependent variables. In this case as well, the potential effect of data clustering on the CAARMS' diagnosis was preliminarily assessed and entered into the ROC regression model (Alonzo & Pepe, 2002; Cai, 2004; Faraggi, 2003; Pepe, 1998; Rodriguez-Alvarez et al., 2011). For both scales, the overall accuracy and validity of the iPQ-16 were estimated using the area under the ROC curve (AUC; based on 5000 stratified bootstrap resamples) and interpreted according to Swets's benchmarks (Swets, 1998; Zweig & Campbell, 1993). Additionally, Sensitivity (Se) and Specificity (Sp) were computed for the selected cut-off point (Pepe, 2003; Zhou et al., 2002).

3 | RESULTS

3.1 | Preliminary analysis: Effect of data clustering

No items from the 'A/P-scale' were associated with an ICC or MOR higher than the recommendedthresholds (Table 1). None of the items of the 'I-scale' revealed an ICC or a MRR higher than the recommend thresholds (Table 2). These results indicate no clustering effect of the mental health department and/or the high-school/university on items responses.

3.2 | Structural validity

3.2.1 | Absence/presence scale ('a/P-scale')

A single-factor first-order model showed an acceptable fit to the A/P-scale data for the two samples combined together (Table 3). Despite the fact that the χ^2 was significant [χ^2 (104) = 186.460; p < 0.001 and χ^2 (104) = 258.906; p < 0.001], the RMSEA [0.032; 90% CI 0.025–0.032; p(<0.05) = 1 and 0.044; 90% CI 0.038–0.051; p(<0.05) = 0.911], the CFI (CFI = 0.989 and CFI = 0.985), and the χ^2/df (χ^2/df = 1.793 and χ^2/df = 2.489) were indicative of a good model fit.

With respect to UHR patients (M = 5.38, SD = 4.22), although the χ^2 tests yielded statistical significance [χ^2 (104) = 186.735; p < 0.001 and χ^2 (104) = 249.684; p < 0.001], the RMSEA [RMSEA = 0.042; 90% CI 0.032–0.052; p(<0.05) = 0.901 and RMSEA = 0.056; 90% CI 0.048–0.065; p(<0.05) = 0.113], the CFI (0.988 and 0.981), and the χ^2/df (1.795 and 2.400) suggested a good model fit. For individuals enrolled from the general population (M = 4.20, SD = 3.10), the χ^2 test was non-statistically significant for the A/P scale [χ^2 (104) = 102.912; p = 0.512 ns] and statistically significant for the I-scale [χ^2 (104) = 135.503; p < 0.001]. However, the RMSEA [RMSEA = 0.000; 90% CI 0.000–0.028; p(<0.05) = 1 and RMSEA = 0.031; 90% CI 0.013–0.045; p(<0.05) = 0.991], the CFI (1 and 0.989), and the χ^2/df (1.712 and 1.303) indicated a good model fit in both cases.

3.2.2 | Intensity scale ('I-scale')

A single-factor first-order model demonstrated an acceptable fit to the I-scale data for the two combined samples (Table 3). Despite the

TABLE 1 Descriptive statistics (%), ICC and MOR, and item discriminant power for the PQ-16 *presence/absence* scale (response scale: 0–1).

	Descriptive statistics		Multilevel clustering effect		Item discriminant power		
	E (%)	¬*E* (%)	ICC	MOR	*z*	IDA	*h*
Item#1	60.8	39.2	0.021	1.362	13.595	0.706	1.657
Item#2	46.0	54.0	0.041	1.544	13.993	0.742	1.671
Item#3	81.1	18.9	0.011	1.246	9.595	0.421	1.223
Item#4	73.3	26.7	0.019	1.345	13.680	0.693	1.781
Item#5	71.6	28.4	0.024	1.395	13.989	0.710	1.868
Item#6	88.7	11.3	0.083	1.888	7.953	0.307	1.037
Item#7	60.5	39.5	0.009	1.226	13.452	0.694	1.673
Item#8	86.2	13.8	0.035	1.497	9.851	0.432	1.296
Item#9	65.8	34.2	0.031	1.454	15.110	0.784	1.992
Item#10	68.1	31.9	0.052	1.637	12.125	0.609	1.474
Item#11	53.7	46.3	0.012	1.265	16.273	0.859	2.119
Item#12	72.8	27.2	0.009	1.219	11.814	0.580	1.477
Item#13	85.0	15.0	0.031	1.458	10.273	0.455	1.480
Item#14	52.0	48.0	0.010	1.230	15.095	0.797	1.869
Item#15	77.9	22.1	0.025	1.405	12.054	0.585	1.604
Item#16	71.1	28.9	0.027	1.421	12.530	0.614	1.663

Note: *E* = % of item endorsement; ¬*E* = % of item non-endorsement; ICC = intraclass correlation coefficient based on Snijders and Bosker (1999) formula and binomial distribution; MOR = median odds ratio based on Merlo et al. (2006) formula; IDP = item discriminant power; *z* = z-statistic; IDA = Ebel's item discriminant ability; *h* = Ebel's effect size of item discriminant ability.

TABLE 2 Descriptive statistics, ICC and MOR, and item discriminant power for the PQ-16 *intensity* scale (response scale: 0–3).

	Descriptive statistics				Multilevel clustering effect		Item discriminant power		
	Mean	SD	Skw.	*K*	ICC	MRR	*t*	*d*	$r_{(it\text{-}tot)}$
Item#1i	0.55	0.844	1.353	0.745	0.021	1.366	−16.435	1.695	0.504
Item#2i	0.52	0.761	1.350	1.085	0.011	1.255	−14.147	1.452	0.435
Item#3i	0.15	0.468	3.607	13.728	0.021	1.358	−7.488	0.774	0.314
Item#4i	0.34	0.693	2.164	4.193	0.022	1.375	−13.265	1.371	0.557
Item#5i	0.43	0.827	1.880	2.454	0.011	1.244	−14.532	1.501	0.578
Item#6i	0.13	0.431	4.029	18.569	0.074	1.815	−6.819	0.705	0.373
Item#7i	0.64	0.931	1.198	0.171	0.009	1.221	−16.318	1.682	0.407
Item#8i	0.18	0.576	3.500	12.138	0.037	1.508	−8.433	0.872	0.471
Item#9i	0.48	0.858	1.735	1.931	0.033	1.474	−16.063	1.660	0.602
Item#10i	0.26	0.606	2.550	6.533	0.038	1.516	−10.011	1.034	0.428
Item#11i	0.73	0.974	1.035	−0.213	0.019	1.341	−23.421	2.417	0.615
Item#12i	0.26	0.581	2.346	5.226	0.003	1.129	−10.701	1.105	0.422
Item#13i	0.21	0.614	3.260	10.390	0.034	1.481	−9.366	0.969	0.486
Item#14i	0.78	1.010	0.954	−0.423	0.005	1.167	−20.544	2.119	0.531
Item#15i	0.30	0.677	2.344	4.795	0.019	1.339	−12.798	1.323	0.564
Item#16i	0.37	0.771	2.126	3.621	0.012	1.267	−12.446	1.304	0.467

Note: SD = standard deviation; Skw. = skewness; K; ICC = intraclass correlation coefficient based on Snijders and Bosker (1999) formula and probit distribution; MRR = median rate ratio (Austin et al., 2018); IDP = item discriminant power; *t* = independent sample *t*-test; *d* = effect size – Cohen's d; *r*(it-tot) = item-total correlation;

statistically significant S-Bχ^2 [S-Bχ^2 (104) = 258.906; $p < 0.001$], the RMSEA [RMSEA = 0.044; 90% CI 0.038–0.051; p(RMSEA <0.05) = 0.911], the CFI (CFI = 0.985), and the χ^2/df (χ^2/df = 2.489), indicated a good model fit. For the UHR patient sample (M = 7.46, SD = 7.29), although the S-Bχ^2 was statistically significant [S-Bχ^2 (104) = 249.684; $p < 0.001$], other fit indices, including the RMSEA

TABLE 3 Confirmatory factor analysis for both the PQ-16 *absence/presence* scale (scaled: 0–1) and the PQ-16 *intensity* scale (scaled: 0–3).

	PQ-16 *presence/absence* scale						PQ-16 *intensity* scale					
	Overall sample		UHR patients		General population		Overall sample		UHR patients		General population	
	λ	R^2	λ	R^2	λ	R^2	λ	R^2	λ	R^2	λ	R^2
Item#1	0.577	0.333	0.575	0.330	0.551	0.304	0.613	0.376	0.592	0.350	0.612	0.375
Item#2	0.585	0.342	0.743	0.553	0.458	0.210	0.549	0.302	0.600	0.360	0.481	0.231
Item#3	0.501	0.251	0.712	0.507	0.185	0.034	0.546	0.298	0.605	0.366	0.323	0.104
Item#4	0.722	0.522	0.793	0.628	0.541	0.293	0.731	0.534	0.772	0.596	0.604	0.364
Item#5	0.728	0.531	0.760	0.577	0.645	0.416	0.728	0.530	0.750	0.563	0.676	0.457
Item#6	0.665	0.442	0.631	0.398	0.693	0.481	0.643	0.413	0.595	0.354	0.727	0.529
Item#7	0.552	0.304	0.542	0.294	0.560	0.313	0.538	0.290	0.521	0.272	0.538	0.289
Item#8	0.766	0.587	0.783	0.613	0.632	0.399	0.749	0.561	0.753	0.567	0.648	0.420
Item#9	0.767	0.588	0.773	0.598	0.737	0.544	0.755	0.569	0.723	0.523	0.777	0.604
Item#10	0.588	0.346	0.670	0.449	0.533	0.284	0.611	0.373	0.591	0.350	0.598	0.358
Item#11	0.700	0.491	0.741	0.548	0.622	0.387	0.752	0.565	0.758	0.575	0.733	0.537
Item#12	0.610	0.373	0.710	0.503	0.470	0.221	0.608	0.369	0.652	0.425	0.510	0.260
Item#13	0.758	0.574	0.770	0.592	0.664	0.441	0.758	0.575	0.785	0.616	0.644	0.415
Item#14	0.621	0.386	0.647	0.418	0.612	0.374	0.648	0.420	0.612	0.374	0.755	0.570
Item#15	0.732	0.536	0.726	0.527	0.710	0.504	0.767	0.589	0.753	0.566	0.771	0.594
Item#16	0.599	0.358	0.597	0.356	0.645	0.417	0.621	0.386	0.588	0.346	0.686	0.471

Note: All factor loadings (λ) and explained variances (R^2) are statistically significant at $p < .001$.

[RMSEA = 0.056; 90% CI 0.048–0.065; p(RMSEA <0.05) = 0.113], the CFI (CFI = 0.981), and the χ^2/df ($\chi^2/df = 2.400$), suggested a good model fit. For the sample of individuals enrolled from the general population ($M = 4.73$, SD = 5.41), despite the statistically significant S-Bχ^2 [S-Bχ^2 (104) = 135.503; $p < 0.001$], other fit indices, including the RMSEA [RMSEA = 0.031; 90% CI 0.013–0.045; p(RMSEA <0.05) = 0.991], the CFI (CFI = 0.989), and the χ^2/df ($\chi^2/df = 1.303$), suggested a good model fit.

3.3 | Measurement invariance

3.3.1 | Absence/presence scale ('a/P scale')

Configural invariance: The model exhibited good fit indices ($\chi^2 = 289.647$, $p < .001$; RMSEA = 0.032; 90% CI 0.023–0.041; p(RMSEA <0.05) = 1, CFI = 0.990, χ^2/df = 1.392), indicating that the factorial structure was consistent across groups.

Metric + Scalar Invariance: This model also demonstrated a good fit ($\chi^2 = 358.792$, $p < 0.001$; RMSEA = 0.040; 90% CI 0.032–0.048; p(RMSEA <0.05) = 0.985, CFI = 0.983, χ^2/df = 1.616). Although there was a slight decrease in fit indices compared to the configural model, it was non-significant (DIFTEST = 69.145, $p < 0.001$; ΔRMSEA = 0.008; ΔCFI = −0.007), indicating that items were equally related to the latent factor between the groups and had the same expected item response at the same absolute level of the trait.

3.3.2 | Intensity scale ('I-scale')

Configural invariance: The model exhibited good fit indices ($\chi^2 = 385.187$, $p < 0.001$; RMSEA = 0.048; 90% CI 0.040–0.055; p(RMSEA <0.05) = 0.703, CFI = 0.983, χ^2/df = 1.851), indicating that the factorial structure was consistent across groups.

Metric Invariance: This model still provided a good fit ($\chi^2 = 485.896$, $p < 0.001$; RMSEA = 0.056; 90% CI 0.049–0.063; p(RMSEA <0.05) = 0.075, CFI = 0.975, χ^2/df = 2.178). Compared to the configural model, there was a non-significant decrease in two out of three fit indices (DIFTEST = 100.71, $p < 0.001$; ΔRMSEA = 0.008; ΔCFI = −0.008), indicating that items were equally related to the latent factor between the groups.

Scalar invariance: The scalar invariance model also fit the data well ($\chi^2 = 517.258$, $p < 0.001$; RMSEA = 0.052; 90% CI 0.046–0.059; p(RMSEA <0.05) = 0.265, CFI = 0.974, χ^2/df = 2.036). Compared to the configural model, there was a non-significant decrease in all three fit indices (DIFTEST = 31.362, $p = 0.448$; ΔRMSEA = 0.003; ΔCFI = −0.001), suggesting that items had the same expected item response at the same absolute level of the trait (Table 4).

3.4 | Psychometric properties

The IDP analysis showed that each of the 16 items discriminated well between subjects with the absence/presence of prodromal symptoms as well as low or high symptom distress. For the 'A/P-scale', the

TABLE 4 Measurement invariance analysis for both the PQ-16 *absence/presence* scale and the PQ-16 *intensity* scale.

	χ^2 (*df*)	RMSEA	CFI	$\Delta\chi^2$	*p*-Value	|ΔRMSEA|	|ΔCFI|
A/P-scale							
UHR patients	186.735 (104)	0.042	0.988				
General population	102.912 (104)	0.000	1.000				
Configural invariance	289.647 (208)	0.032	0.990				
Metric + Scalar invariance	358.792 (222)	0.040	0.983	69.145	<.001	0.008	0.007
I-scale							
UHR patients	249.684 (104)	0.056	0.981				
General population	135.503 (104)	0.031	0.989				
Configural invariance	385.187 (208)	0.048	0.983				
Metric invariance	485.896 (223)	0.056	0.975	100.71	<.001	0.008	0.008
Scalar invariance	517.258 (254)	0.052	0.974	31.362	.448	0.003	0.001

Abbreviations: χ^2, chi-square test; df, degree of freedoms; Δ, differences between indices; RMSEA, root mean square error of approximation; CFI, comparative fit index.

TABLE 5 Correlation between scales.

		M	SD	1	2	3	4	5	6	7
1	PQ-16, 'A/P-scale'	4.85	3.829	-						
2	PQ-16, 'I-scale'	6.33	6.713	0.858**	-					
3	SOFAS	58.94	14.024	−0.309**	−0.275**	-				
4	HONOS	10.32	6.574	0.170**	0.237**	−0.486**	-			
5	HONOSCA	12.68	7.810	0.283**	0.241**	−0.631**	0.975**	-		
6	GAF	54.50	13.296	−0.167**	−0.214**	0.883**	−0.432**	−0.552**	-	
7	GHQ	11.87	8.553	0.312**	0.356**	−0.308**	0.378**	0.202**	−0.164*	-

*p < .050; **p < .001.

discrimination parameter *IDA* ranged from 0.307 (item#6; 'When I look at a person, or look at myself in a mirror, I have seen the face change right before my eyes') to 0.859 (item#11; 'Sometimes I have felt that I'm not in control of my own ideas or thoughts'), with an associated effect size that ranged from 1.037 to 2.199 – Table 1. For the 'I-scale', the discrimination parameter *t* ranged from |6.819| (item#6; 'When I look at a person, or look at myself in a mirror, I have seen the face change right before my eyes') to |23.421| (item#11; 'Sometimes I have felt that I'm not in control of my own ideas or thoughts'), with an associated effect size that ranged from 0.705 to 2.417 – Table 2. The adjusted item-total correlation also revealed a moderate-to-strong association between each item and the total score. Reliability analysis revealed good results: for the 'A/P-scale', ω was 0.940, and for the 'I-scale' ω was 0.950. Additionally, test–retest reliability showed excellent results: the two-way mixed ICC was 0.965 (95% CI 0.934–0.982) for the 'A/P-scale' and 0.957 (95% CI 0.919–0.977) for the 'I-scale', respectively. Small-to-moderate correlations were found between the iPQ-16 scales and SOFAS, HoNOS, HoNOSCA, GAF, and GHQ (Table 5).

Considering the CAARMS diagnosis, UHR patients ($M = 7.44$, SD = 3.58) showed statistically significantly higher values in the number of endorsed symptoms ('A/P-scale') compared with non-UHR patients ($M = 3.62$, SD = 3.59): $U = 963$, $z = -5.091$ $p < 0.001$, $g = -1.064$. When examining the I-scale, UHR patients ($M = 13.65$, SD = 8.19) exhibited a significantly higher intensity of endorsed symptoms in comparison to non-UHR patients ($M = 5.16$, SD = 6.14): $t = -5.640$, $p < 0.001$, $g = 1.284$.

3.5 | Accuracy of the PQ-16 as a screening/diagnostic tool

Considering that the external criterion variable should be influenced by the 2nd level of data clustering (mental health department), ICC for binary data and MOR were performed. The results revealed no effect of the data clustering: ICC = 0.027, MOR = 1.420. Thus, mental health departments were treated as a covariate. An ordinary ROC analysis was performed, and the ROC-regression analysis revealed no influence of the covariate on the external criterion variable. Unstandardized regression coefficients ranged from 0.813 (95% CI: −1.525 to 3.138; $p = 0.498$), to 1.916 (95% CI: −0.483 to 4.285; $p = 0.118$). The 'A/P-scale' of the iPQ-16 demonstrated good accuracy in

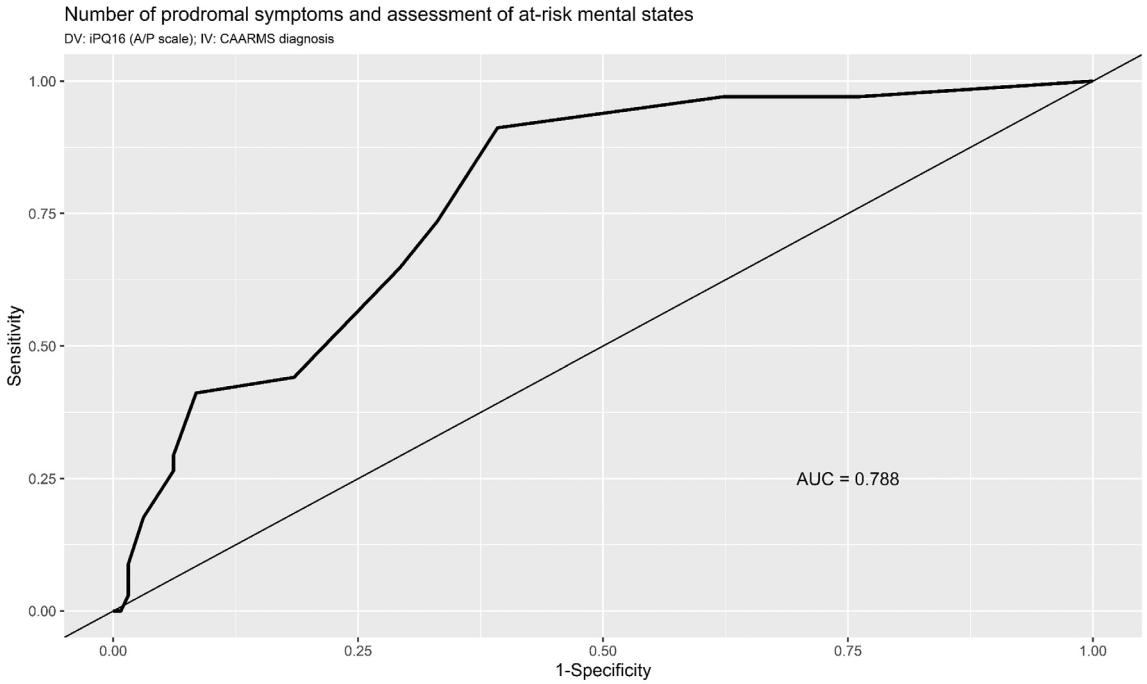

FIGURE 1 ROC curve analysis.

discriminating between non-UHR and UHR patients based on the CAARMS diagnosis: AUC = 0.788; 95% CI = 0.708–0.867 (Figure 1). In addition, the cut-off value of 4 showed the best discriminating properties, with a sensitivity of 0.909 and a specificity of 0.611. The traditional cut-off of 6 showed a sensitivity of 0.667 and specificity of 0.714.

4 | DISCUSSION

In our study, we conducted a comprehensive psychometric evaluation of the iPQ-16 using data from help-seeking outpatients in multiple mental health departments and individuals from the general population in Italy. Overall, our findings support the iPQ-16 as a valid and reliable screening tool for UHR symptoms with a single-factor first-order (unidimensional) structure, allowing for an easy comparison between UHR outpatients and individuals from the general population (scalar measurement invariance).

One notable novel contribution of our study is the examination of the factorial structure of the iPQ-16, which has not been previously tested. Our confirmatory factor analysis revealed a good fit for a single-factor first-order model, indicating that all 16 items of the iPQ-16 are aligned with the general dimension of the scale in both samples. This supports the use of a total score derived from the simple sum of the 16 items as a representative measure of UHR symptoms. Furthermore, our analysis showed that a single-factor first-order model has an acceptable fit for both the 'A/P scale' and the 'I-scale' of the iPQ-16. This is particularly significant as these scoring scales have been widely used (Kline et al., 2014), and the inclusion of the 'I-scale' enhances the identification of UHR individuals and improves the accuracy of the PQ-16, particularly in terms of specificity (Savill et al., 2018).

Even more importantly, we verified the measurement invariance as a precondition for comparing scores and indicators across both help-seeking and non-help-seeking samples. The analysis explored potential construct differences between the two samples based on item metric properties for both scales ('A/P-scale' and 'I-Scale') (Brown, 2015; Meredith, 1993;). For both scales ('A/P-scale' and 'I-Scale'), strong invariance was achieved. Furthermore, for the first time, PQ-16 invariance was tested between two samples with clearly different psychological/psychiatric characteristics. Participants in the help-seeking and non-help-seeking samples interpreted the iPQ-16 items equivalently. Therefore, the PQ-16 scoring procedure can be applied equivalently in the help-seeking population and in the general population, using the same scoring procedure. The significance of these findings lies in the potential use of the iPQ-16 as a screening tool in non-help-seeking samples, such as schools, to identify individuals at risk. However, caution is advised when making comparisons, as our study revealed differences between help-seeking patients and the general population (Manzoni et al., 2020; Meredith, 1993; van de Schoot et al., 2012). Notably, certain items (e.g., #4, #8, #13) were more prevalent in help-seeking subjects, reflecting higher factor loadings, while items (e.g., #6, #9, #15) were more representative of the general population. Items #4, #8, and #13 specifically target subtle hallucinatory experiences and are more relevant to our UHR target group. Conversely, items #6, #9, and #15 have a less specific sensorial connotation and may be prone to misinterpretation. This variability in item representation aligns with Savill et al.'s insights, emphasizing the

need for proper item interpretation in different settings and underscoring the importance of considering contextual factors in the application of psychosis risk identification tools like the iPQ-16 (Savill et al., 2018).

We also conducted an investigation into the classical psychometric properties and convergent validity analyses of the iPQ-16. The instrument demonstrated high internal consistency for both scales, consistent with existing literature for the A/P scoring system (Azzali et al., 2018; Chen, Wang, Heeramun-Aubeeluck, et al., 2014; de Jong et al., 2018; Ising et al., 2012; Pelizza et al., 2018). Notably, this result is novel regarding the 'I-scale'. Furthermore, the iPQ-16 exhibited excellent one-month test-retest reliability. We observed significant positive correlations between the iPQ-16 and other global scales assessing psychological distress, global functioning, and global psychopathology, as well as with the CAARMS, aligning with previous validation studies (Azzali et al., 2018; Pelizza et al., 2018).Lastly, the results from the ROC analyses highlighted the effectiveness of the iPQ-16, particularly when utilizing the 'A/P scale', as a robust screening tool for identifying high-risk states. The iPQ-16 demonstrated good accuracy (AUC = 0.788), moderate specificity (0.611), and high sensitivity (0.909) in distinguishing between patients with UHR for psychosis and those without UHR for psychosis, with an optimal cut-off value of 4. This cut-off value differs from the original PQ-16, which had a cut-off of ≥6. Our proposed cut-off of ≥4 significantly enhanced sensitivity to 91%, albeit at the expenses of specificity, reduced to 61%. In contrast, a previous Italian single-centre study reported a sensitivity of 69% and specificity of 83% with an optimal cut-off of ≥5. These discrepancies in findings may be attributed to differences in selection procedures (e.g., screening, triage) and/or sample sizes. Savill et al.'s systematic review underscored the variability of thresholds when using the Prodromal Questionnaire in different settings, reminding us of the delicate balance needed between sensitivity and specificity in such tools (Savill et al., 2018). The iPQ-16 is more effective in identifying individuals who may have UHR/psychosis (ruling in) rather than ruling out those who do not have UHR/psychosis. Therefore, it can be used as part of a comprehensive screening approach, where individuals who screen positive can undergo further assessment using a more comprehensive clinician-rated scale like the CAARMS (Savill et al., 2018). This two-step process allows for a more accurate identification of individuals at risk and ensures that appropriate interventions are provided. In conclusion, adhering to current scientific guidelines, our methodology and statistical analysis enabled us to the examine the iPQ-16's internal structure, measurement invariance, and identify an optimal clinical cut-off. Our findings collectively establish the iPQ-16 as a valid and robust screening tool for identifying UHR individuals. However, it is important to acknowledge several limitations in our study. Firstly, the multi-centric nature of the study and the unequal sample sizes across centers may have introduced some degree of heterogeneity in the data. To address this, we conducted a preliminary analysis to assess any potential clustering effect of recruitment centers, effectively controlling this bias. Secondly, the test–retest validity analysis was based on a smaller sub-sample of the help-seeking population. While current guidelines suggest that a sample size of 40 is sufficient for such analyses (Tabachnick & Fidell, 2014), the relatively smaller sample size could be considered a limitation. Thirdly, the non-help-seeking population only completed the iPQ-16 and not the CAARMS. While our findings suggest that the iPQ-16 could potentially be used in the general population, further investigations are needed to evaluate its cost-effectiveness and determine the specific cut-off values for this particular population.

Another limitation to consider is the potential difference in functioning between UHR subjects and presumed healthy subjects recruited from the general population. Typically, UHR subjects may exhibit poorer functioning compared to presumed healthy individuals (Salazar de Pablo et al., 2021). However, our primary aim was to recruit presumed healthy subjects who were as comparable as possible in terms of age, sex, and socio-demographic status to the UHR subjects.In conclusion, our study highlights the potential of the iPQ-16 as a valuable tool for screening individuals at risk of UHR/psychosis. However, it is crucial to consider the limitations of the scale, particularly in terms of its concurrent validity.

5 | CONCLUSIONS

Based on our findings, the iPQ-16 demonstrates its effectiveness as a valuable tool for the initial selection and screening of individuals at ultra-high risk (UHR) for psychosis, both within the help-seeking population and the general population. It also proves to be a valid and reliable instrument for research purposes related to UHR mental states. Our study extends the insights from Savill et al. (2018), highlighting the importance of adaptability and context-specific application in psychosis risk assessments. The iPQ-16 has shown potential as a valuable tool for screening individuals at risk of UHR/psychosis, demonstrating the value of the nuanced application of risk identification and threshold adjustment (Savill et al., 2018).

Considering these results, we strongly recommend promoting and utilizing the iPQ-16 as a validated assessment tool for the detection of UHR subjects within the context of Italy's National Health Service. By incorporating the iPQ-16 into clinical practice, healthcare professionals can enhance their ability to identify individuals at risk for psychosis and provide appropriate interventions and support. Furthermore, the iPQ-16 can contribute to a more systematic and efficient identification of UHR individuals, leading to improved early intervention and potentially influencing the trajectory of their mental health outcomes.

Further research should address the limitations of our study and explore the utility of the iPQ-16 in diverse populations and settings. It is important to continue validating the scale and establishing its effectiveness in different contexts to enhance its clinical utility. By doing so, we can improve the early detection and intervention for individuals at risk of UHR/psychosis and ultimately contribute to better mental health outcomes.

ACKNOWLEDGEMENTS

Team involved in CCM2013 Project: M. Luisa Abbinante, Evita Barbera, Alice Bizzozero, Serena Boni, Agnese Canali, Micaela Corio,

Elena Criconia, Francesaca Fauci, Maria Teresa Gris, Monica Maggi, Simonetta Oriani, Silvia Pinotti, Katia Prato, Chiara Primati, Maria Teresa Rivetti, Maria Teresa Rossi, Valentina Scavelli, Monica Soffientini, Daniele Tuso, Zaira Xodo (ASST Rhodense, Garbagnate Milanese); Barbara D'Avanzo, Alberto Parabiaghi (Istituto di Ricerche Farmacologiche Mario Negri); Aglaia Banis Mariano Bassi, Elisa Buratti, Patrizia Conti, Maria Meliante, Anita Montanari, Emiliano Monzani, Mauro Emilio Percudani, Piero Rossi (ASST Grande Ospedale Metropolitano 'Niguarda', Miano); Irene Anzani, Maria Teresa Asti, Giuseppina Badia, Giuseppe Cardamone, Nadia Magnani, Stefano Milano, Silvia Sordini, Annalisa Spampani (Azienda USL 9 Grosseto); Alberto Corbetta, Alessandra Ferrara, Antonio Lora, Ottaviano Martinelli, Donatella Puccia, Vittorio Rigamonti, Elisa Stucchi, Marina Zabarella, Sabrina Zanetti (ASST di Lecco); Alessandra Barni, Emanuela Butteri, Deborah Fiocca, Ornella Giardina, Teodoro Maranesi, Claudio Monaci, Silvia Olivieri, Antonella Perri, Manuela Pintori, Caterina Roncoroni, Michele Santi, Annapatrizia Stipcovich, Raffaella Zanaschi (ASST Fatebenefratelli-Sacco, Milano); Maria Pia Angelatos, Paola Arnuzzo, Cristina Bonzani, Selene Cammarata, Miriam Cantatore, Patrizia Capurro, Michela Davi, Margherita Dolcino, Lorenza Ferrari, Marta Germoglio, Laura Grondona, Maria Guelfi, Lidia Lagomarsino, Alessandro Lanza, Alessandra Lombardo, Eleonora Lusetti, Patrizia Marinelli, Claudia Masala, Emanuela Mirrione, Diana Monachesi, Lucia Pacini, Sara Patti, Elisa Peloso, Laura Penco, Maria Franca Petrigni, Giulia Piccinini, Sara Pignatelli, Valeria Puppo, Cristina Rambelli, Daniela Ratti, Alessandra Ravecca, Rossana Roveda, Rita Schenone, Lucia Sciarretta, Gianluca Serafini, Irene Serio, Simonetta Sommariva, Lucia Tacconi, Alessandra Tauriello, Ludovica Tognoni, Fabrizia, Tomasi, Simona Traverso, Giovanna Trompetto, Serena Unia, Marco Vaggi, Cristina Venturino, Alessandra Zaccone, Monica Zambonini, Franca Zanelli (ASL 3 Genova); Davide Maestri (ASST Santi Paolo e Carlo, Milano). Open access funding provided by BIBLIOSAN.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because due to privacy restrictions, data were available from the corresponding author on a reasonable request.

ORCID

Alberto Parabiaghi https://orcid.org/0000-0002-2898-5134
Alessandro Alberto Rossi https://orcid.org/0000-0001-7000-5999
Anna Castelnovo https://orcid.org/0000-0002-8875-9657
Lorenzo del Fabro https://orcid.org/0000-0003-1452-9688
Stefania Mannarini https://orcid.org/0000-0002-8446-785X

REFERENCES

Aas, I. H. M. (2011). Guidelines for rating global assessment of functioning (GAF). *Annals of General Psychiatry*, *10*(1), 2. https://doi.org/10.1186/1744-859x-10-2

Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management*, *39*(6), 1490–1528. https://doi.org/10.1177/0149206313478188

Alonzo, T. A., & Pepe, M. S. (2002). Distribution-free ROC analysis using binary regression techniques. *Biostatistics*, *3*(3), 421–432. https://doi.org/10.1093/biostatistics/3.3.421

Anderson, C. J., Kim, J.-S., & Keller, B. (2014). Multilevel modeling of categorical response variables. In *Handbook of international large-scale assessment* (pp. 481–519). Taylor & Francis Group.

Austin, P. C., & Merlo, J. (2017). Intermediate and advanced topics in multilevel logistic regression analysis. *Statistics in Medicine*, *36*, 3257–3277.

Austin, P. C., Stryhn, H., Leckie, G., & Merlo, J. (2018). Measures of clustering and heterogeneity in multilevel Poisson regression analyses of rates/count data. *Statistics in Medicine*, *37*(4), 572–589. https://doi.org/10.1002/sim.7532

Azzali, S., Pelizza, L., Paterlini, F., Garlassi, S., Scazza, I., Chiri, L. R., Poletti, M., Fontana, F., Favazzo, R., Barbanti, S. V., Fabiani, M., Pensieri, L., Cioncolini, L., Semrov, E., & Raballo, A. (2018). Reliability of the Italian version of the 16-item prodromal questionnaire (iPQ-16) or psychosis risk screening in a young help-seeking community sample. *Journal of Psychopathology*, *24*, 16–23.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815–824. https://doi.org/10.1016/j.paid.2006.09.018

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine (Phila PA 1976)*, *25*(24), 3186–3191.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. http://www.ncbi.nlm.nih.gov/pubmed/2320703

Berchtold, A. (2016). Test–retest: Agreement or reliability? *Methodological Innovations*, *9*, 2059799116672875. https://doi.org/10.1177/2059799116672875

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research*, *24*(4), 445–455. https://doi.org/10.1207/s15327906mbr2404_4

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Cai, T. (2004). Semi-parametric ROC regression analysis with placement values [research support, U.S. Gov't, P.H.S.]. *Biostatistics*, *5*(1), 45–60. https://doi.org/10.1093/biostatistics/5.1.45

Chen, F., Wang, L., Heeramun-Aubeeluck, A., Wang, J., Shi, J., Yuan, J., & Zhao, X. (2014). Identification and characterization of college students with attenuated psychosis syndrome in China. *Psychiatry Research*, *216*(3), 346–350. https://doi.org/10.1016/j.psychres.2014.01.051

Chen, F., Wang, L., Wang, J., Heeramun-Aubeeluck, A., Yuan, J., & Zhao, X. (2016). Applicability of the Chinese version of the 16-item prodromal questionnaire (CPQ-16) for identifying attenuated psychosis syndrome in a college population. *Early Intervention in Psychiatry*, *10*(4), 308–315. https://doi.org/10.1111/eip.12173

Chen, F., Wang, L., & Zhao, X. (2014). Validity and reliability of the 16-item version of the prodromal questionnaire in screening psychosis risk of help-seekers. *Chinese Mental Health Journal*, *28*, 667–673.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Chiorri, C. (2011). *Teoria e tecnica psicometrica*. Costruire un test psicologico.

Christensen, R. H. B. (2019). Ordinal—Regression Models for Ordinal Data. In (Version 2019.4-25). http://www.cran.r-project.org/package=ordinal

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

D'Avanzo, B., Lovaglio, P., Parabiaghi, A., Conti, P., Frigerio, A., Molteni, M., Monzani, E., Rossi, P. E., & Barbato, A. (2018). Health of the nation outcome scales for children and adolescents (HoNOSCA): Psychometric properties of the Italian version. *Children and Youth Services Review*, *94*, 340–346. https://doi.org/10.1016/j.childyouth.2018.10.005

de Jong, Y., Mulder, C. L., Boon, A. E., Deen, M., van't Hof, M., & van der Gaag, M. (2018). Screening for psychosis risk among adolescents in child and adolescent mental health services: A description of the first step with the 16-item version of the prodromal questionnaire (PQ-16). *Early Intervention in Psychiatry*, *12*(4), 669–676. https://doi.org/10.1111/eip.12362

de Vet, H. C., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology*, *59*(10), 1033–1039. https://doi.org/10.1016/j.jclinepi.2005.10.015

Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *The Leadership Quarterly*, *16*(1), 149–167. https://doi.org/10.1016/j.leaqua.2004.09.009

Ebel, R. L. (1965). *Measuring educational achievement*. Prentice-Hall.

Faraggi, D. (2003). Adjusting receiver operating characteristic curves and related indices for covariates. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *52*(2), 179–192. https://doi.org/10.1111/1467-9884.00350

Fusar-Poli, P., Borgwardt, S., Bechdolf, A., Addington, J., Riecher-Rössler, A., Schultze-Lutter, F., Keshavan, M., Wood, S., Ruhrmann, S., Seidman, L. J., Valmaggia, L., Cannon, T., Velthorst, E., De Haan, L., Cornblatt, B., Bonoldi, I., Birchwood, M., McGlashan, T., Carpenter, W., … Yung, A. (2013). The psychosis high-risk state: A comprehensive state-of-the-art review. *JAMA Psychiatry*, *70*(1), 107–120. https://doi.org/10.1001/jamapsychiatry.2013.269

Fusar-Poli, P., Hobson, R., Raduelli, M., & Balottin, U. (2012). Reliability and validity of the comprehensive assessment of the at risk mental state, Italian version (CAARMS-I). *Current Pharmaceutical Design*, *18*(4), 386–391. https://doi.org/10.2174/138161212799316118

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*(1), 72–91. https://doi.org/10.1037/a0032138

Goldberg, D. P. (1972). *The detection of psychiatric illness by questionnaire*. Oxford University Press.

Goldman, H. H., Skodol, A. E., & Lave, T. R. (1992). Revising axis V for DSM-IV: A review of measures of social functioning. *American Journal of Psychiatry*, *149*(9), 1148–1156.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics*, *1*(4), 223–231. https://doi.org/10.1207/S15328031US0104_02

Gowers, S., Bailey-Rogers, S. J., Shore, A., & Levine, W. (2000). The Health of the Nation Outcome Scales for Child & Adolescent Mental Health (HoNOSCA). *Child Psychology and Psychiatry Review*, *5*(2), 50–56. https://doi.org/10.1017/s1360641700002148

Gowers, S. G., Harrington, R. C., Whitton, A., Lelliott, P., Beevor, A., Wing, J., & Jezzard, R. (1999). Brief scale for measuring the outcomes of emotional and behavioural disorders in children: Health of the nation outcome scales for children and adolescents (HoNOSCA). *British Journal of Psychiatry*, *174*(5), 413–416. https://doi.org/10.1192/bjp.174.5.413

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, *74*(1), 155–167. https://doi.org/10.1007/s11336-008-9099-3

Grimm, K. J., Ram, N., & Estabrook, R. (2017). *Growth modeling: Structural equation and multilevel modeling approaches*. Guilford Publications.

Guillemin, F., Bombardier, C., & Beaton, D. E. (1993). Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. *Journal of Clinical Epidemiology*, *46*(12), 1417–1432.

Hall, R. C. W. (1995). Global assessment of functioning: A modified scale. *Psychosomatics*, *36*(3), 267–275. https://doi.org/10.1016/S0033-3182(95)71666-8

Hayes, A. F. (2006). A primer on multilevel modeling. *Human Communication Research*, *32*, 385–410. https://doi.org/10.1111/j.1468-2958.2006.00281.x

Heck, R. H. (2001). Multilevel modeling with SEM. In G. A. Marculides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 109–148). Psychology Press.

Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus*. Routledge.

Hedges, L. V., Hedberg, E. C., & Kuyper, A. M. (2012). The variance of Intraclass correlations in three- and four-level models. *Educational and Psychological Measurement*, *72*(6), 893–909. https://doi.org/10.1177/0013164412445193

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Hirschfeld, G., & Von Brachel, R. (2014). Multiple-group confirmatory factor analysis in R-A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research & Evaluation*, *19*(7), 1–12.

Howell, D. C. (2013). *Statistical methods for psychology*. Wadsworth.

Howie, C., Hanna, D., Shannon, C., Davidson, G., & Mulholland, C. (2022). The structure of the prodromal Questionnaire-16 (PQ-16): Exploratory and confirmatory factor analyses in a general non-help-seeking population sample. *Early Intervention in Psychiatry*, *16*(3), 239–246. https://doi.org/10.1111/eip.13147

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications*. Routledge.

Hoyle, R. H. (2012). *Handbook of strucural equation modeling*. The Guilford Press.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Ising, H. K., Veling, W., Loewy, R. L., Rietveld, M. W., Rietdijk, J., Dragt, S., Klaassen, R. M. C., Nieman, D. H., Wunderink, L., Linszen, D. H., & van der Gaag, M. (2012). The validity of the 16-item version of the prodromal questionnaire (PQ-16) to screen for ultra high risk of developing psychosis in the general help-seeking population. *Schizophrenia Bulletin*, *38*(6), 1288–1296. https://doi.org/10.1093/schbul/sbs068

Jones, S. H., Thornicroft, G., Coffey, M., & Dunn, G. (1995). A brief mental health outcome scale: Reliability and validity of the global assessment of functioning (GAF). *British Journal of Psychiatry*, *166*(5), 654–659. https://doi.org/10.1192/bjp.166.5.654

Kline, E., & Schiffman, J. (2014). Psychosis risk screening: A systematic review. *Schizophrenia Research*, *158*(1), 11–18. https://doi.org/10.1016/j.schres.2014.06.036

Kline, E., Thompson, E., Bussell, K., Pitts, S. C., Reeves, G., & Schiffman, J. (2014). Psychosis-like experiences and distress among adolescents using mental health services. *Schizophrenia Research*, *152*(2), 498–502. https://doi.org/10.1016/j.schres.2013.12.012

Kline, E., Wilson, C., Ereshefsky, S., Denenny, D., Thompson, E., Pitts, S. C., Bussell, K., Reeves, G., & Schiffman, J. (2012). Psychosis risk screening in youth: A validation study of three self-report measures of attenuated psychosis symptoms. *Schizophrenia Research*, *141*(1), 72–77. https://doi.org/10.1016/j.schres.2012.07.022

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford Press.

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. https://doi.org/10.1016/j.jcm.2016.02.012

Kotzalidis, G. D., Solfanelli, A., Piacentino, D., Savoja, V., Fiori Nastro, P., Curto, M., Lindau, J. F., Masillo, A., Brandizzi, M., Fagioli, F., Raballo, A., Gebhardt, E., Preti, A., D'Alema, M., Fucci, M. R., Miletto, R., Andropoli, D., Leccisi, D., Girardi, P., … Schultze-Lutter, F. (2017). The Italian version of the 92-item prodromal questionnaire: Concurrent validity with the SIPS and factor analysis in a sample of 258 outpatients aged 11–36years. *Schizophrenia Research*, *189*, 50–56. https://doi.org/10.1016/j.schres.2017.02.008

Loewy, R. L., Bearden, C. E., Johnson, J. K., Raine, A., & Cannon, T. D. (2005). The prodromal questionnaire (PQ): Preliminary validation of a self-report screening measure for prodromal and psychotic syndromes. *Schizophrenia Research*, *79*(1), 117–125. https://doi.org/10.1016/j.schres.2005.03.007

Lora, A., Bai, G., Bianchi, S., Bolongaro, G., Civenti, G., Erlicher, A., Maresca, G., Monzani, E., Panetta, B., Von Morgen, D., Rossi, F., Torri, V., & Morosini, P. (2001). The Italian version of HoNOS (health of the nation outcome scales), a scale for evaluating the outcomes and the severity in mental health services. *Epidemiologia e Psichiatria Sociale*, *10*(3), 198–204. https://doi.org/10.1017/s1121189x00005339

Manzoni, G. M., Rossi, A., Pietrabissa, G., Mannarini, S., Fabbricatore, M., Imperatori, C., Innamorati, M., Gearhardt, A. N., & Castelnuovo, G. (2020). Structural validity, measurement invariance, reliability and diagnostic accuracy of the Italian version of the Yale food addiction scale 2.0 in patients with severe obesity and the general population. *Eating and Weight Disorders – Studies on Anorexia, Bulimia and Obesity*, *26*, 345–366. https://doi.org/10.1007/s40519-020-00858-y

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543.

Merlo, J., Chaix, B., Ohlsson, H., Beckman, A., Johnell, K., Hjerpe, P., Rastam, L., & Larsen, K. (2006). A brief conceptual tutorial of multilevel analysis in social epidemiology: Using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *Journal of Epidemiological Community Health*, *60*, 290–297. https://doi.org/10.1136/jech.2004.029454

Miller, T. J., McGlashan, T. H., Rosen, J. L., Cadenhead, K., Ventura, J., McFarlane, W., Perkins, D. O., Pearlson, G. D., & Woods, S. W. (2003). Prodromal assessment with the structured interview for prodromal syndromes and the scale of prodromal symptoms: Predictive validity, interrater reliability, and training to reliability. *Schizophrenia Bulletin*, *29*(4), 703–715. https://doi.org/10.1093/oxfordjournals.schbul.a007040

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479–515. https://doi.org/10.1207/s15327906mbr3903_4

Montazeri, A., Harirchi, A. M., Shariati, M., Garmaroudi, G., Ebadi, M., & Fateh, A. (2003). The 12-item general health questionnaire (GHQ-12): Translation and validation study of the Iranian version. *Health and Quality of Life Outcomes*, *1*(1), 66. https://doi.org/10.1186/1477-7525-1-66

Muthén, B. O., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus [Unpublished manuscript]. www.statmodel.com/mplus/examples/webnote.html#web4

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

O'Connell, A. A. (2010). An illustration of multilevel models for ordinal response data. Data and context in statistics education: Towards an evidence-based society. In *8th international conference on teaching statistics (ICOTS8)*, Ljubljana, Slovenia.

Parabiaghi, A., Confalonieri, L., Magnani, N., Lora, A., Butteri, E., Prato, K., Vaggi, M., & Percudani, M. E. (2019). Integrated programs for early recognition of severe mental disorders: Recommendations from an Italian multicenter project [perspective]. *Frontiers in Psychiatry*, *10*(844), 1–6. https://doi.org/10.3389/fpsyt.2019.00844

Pelizza, L., Azzali, S., Paterlini, F., Garlassi, S., Scazza, I., Chiri, L. R., Poletti, M., Pupo, S., & Raballo, A. (2019). Screening for psychosis risk among help-seeking adolescents: Application of the Italian version of the 16-item prodromal questionnaire (iPQ-16) in child and adolescent neuropsychiatry services. *Early Intervention in Psychiatry*, *13*(4), 752–760. https://doi.org/10.1111/eip.12554

Pelizza, L., Azzali, S., Paterlini, F., Garlassi, S., Scazza, I., Pupo, S., & Raballo, A. (2018). The Italian version of the 16-item prodromal questionnaire (iPQ-16): Field-test and psychometric features. *Schizophrenia Research*, *199*, 353–360. https://doi.org/10.1016/j.schres.2018.03.023

Pepe, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results [Comparative Study Research Support, U.S. Gov't, P.H.S.]. *Biometrics*, *54*(1), 124–135. http://www.ncbi.nlm.nih.gov/pubmed/9544511

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford Universtiy Press.

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*, 85–112. https://doi.org/10.1016/j.jsp.2009.09.002

Piccinelli, M., & Politi, P. (1993). Struttura fattoriale della versione a 12 domande del General Health Questionnaire in un campione di giovani maschi adulti. *Epidemiologia e Psichiatria Sociale*, *2*(3), 173–181. https://doi.org/10.1017/s1121189x00006990

Pirkis, J. E., Burgess, P. M., Kirk, P. K., Dodson, S., Coombs, T. J., & Williamson, M. K. (2005). A review of the psychometric properties of the health of the nation outcome scales (HoNOS) family of measures [journal article]. *Health and Quality of Life Outcomes*, *3*(1), 76. https://doi.org/10.1186/1477-7525-3-76

R Core Team. (2014). *The R project for statistical computing*. R Core Team www.R-project.org/

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing https://www.R-project.org/

Raballo, A., Semrov, E., Bonner, Y., & Simmons, M. B. (2013). Traduzione e adattamento Italiano della CAARMS (the comprehensive assessment of at risk mental states). Centro Stampa Della Regione Emilia-Romagna.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*(2), 129–140. https://doi.org/10.1080/00223891.2012.725437

Revelle, W. (2018). *Psych: Procedures for personality and psychological research*. Northwestern University.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77. https://doi.org/10.1186/1471-2105-12-77

Rodriguez-Alvarez, M. X., Tahoces, P. G., Cadarso-Suarez, C., & Lado, M. J. (2011). Comparative study of ROC regression techniques. Applications for the computer-aided diagnostic system in breast cancer detection. *Computational Statistics and Data Analysis*, *55*(1), 888–902. https://doi.org/10.1016/j.csda.2010.07.018

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rosseel, Y., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., & Barendse, M. (2015). Package 'lavaan'. http://cran.r-project.org/web/packages/lavaan/lavaan.pdf

Salazar de Pablo, G., Estradé, A., Cutroni, M., Andlauer, O., & Fusar-Poli, P. (2021). Establishing a clinical service to prevent psychosis: What, how and when? Systematic Review. *Translational Psychiatry*, *11*(1), 43. https://doi.org/10.1038/s41398-020-01165-x

Sanagou, M., Wolfe, R., Forbes, A., & Reid, C. M. (2012). Hospital-level associations with 30-day patient mortality after cardiac surgery: A tutorial on the application and interpretation of marginal and multilevel logistic regression. *BMC Medical Research Methodology*, *12*, 28–38. https://doi.org/10.1186/1471-2288-12-28

Savill, M., D'Ambrosio, J., Cannon, T. D., & Loewy, R. L. (2018). Psychosis risk screening in different populations using the prodromal questionnaire: A systematic review. *Early Intervention in Psychiatry*, *12*(1), 3–14. https://doi.org/10.1111/eip.12446

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis*. Sage Publications LTD.

Sommet, N., & Morselli, D. (2017). Keep calm and learn multilevel logistic modeling: A simplified three-step procedure using Stata, R, Mplus, and SPSS. *International Review of Social Psychology*, *30*(1), 203–218. https://doi.org/10.5334/irsp.90

Startup, M., Jackson, M. C., & Bendix, S. (2002). The concurrent validity of the global assessment of functioning (GAF). *British Journal of Clinical Psychology*, *41*(4), 417–422. https://doi.org/10.1348/014466502760387533

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behav Res*, *25*(2), 173–180. https://doi.org/10.1207/s15327906mbr2502_4

Swets, J. A. (1998). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285–1293. https://doi.org/10.1126/science.3287615

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics*. Pearson.

Thomas, S. L., Heck, R. H., & Bauer, K. W. (2005). Weighting and adjusting for design effects in secondary data analyses. *New Directions for Institutional Research*, *127*, 51–72. https://doi.org/10.1002/ir.155

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

van der Gaag, M., Nieman, D. H., Rietdijk, J., Dragt, S., Ising, H. K., Klaassen, R. M. C., Koeter, M., Cuijpers, P., Wunderink, L., & Linszen, D. H. (2012). Cognitive behavioral therapy for subjects at ultrahigh risk for developing psychosis: A randomized controlled clinical trial. *Schizophrenia Bulletin*, *38*(6), 1180–1188. https://doi.org/10.1093/schbul/sbs105

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Wing, J. K., Beevor, A. S., Curtis, R. H., Park, S. G. B., Hadden, J., & Burns, A. (1998). Health of the nation outcome scales (HoNOS): Research and development. *British Journal of Psychiatry*, *172*(1), 11–18. https://doi.org/10.1192/bjp.172.1.11

Wing, J. K., Curtis, R. H., & Beevor, A. S. (1999). Health of the nation outcome scales (HoNOS): Glossary for HoNOS score sheet. *British Journal of Psychiatry*, *174*(5), 432–434. https://doi.org/10.1192/bjp.174.5.432

Yu, C. Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. http://www.statmodel.com/download/Yudissertation.pdf

Yung, A., Phillips, L., & McGorry, P. D. (2004). Treating schizophrenia in the prodromal phase: Back to the future, 1–160. *CRC Press*. https://doi.org/10.3109/9780203501771

Yung, A. R., Yuen, H. P., McGorry, P. D., Phillips, L. J., Kelly, D., Dell'Olio, M., Francey, S. M., Cosgrave, E. M., Killackey, E., Stanford, C., Godfrey, K., & Buckby, J. (2005). Mapping the onset of psychosis: The comprehensive assessment of At-risk mental states [comparative study]. *The Australian and New Zealand Journal of Psychiatry*, *39*(11–12), 964–971. https://doi.org/10.1080/j.1440-1614.2005.01714.x

Zhou, X. H., Obuchowski, N. A., & McClish, D. (2002). *Statistical methods in diagnostic medicine*. Wiley.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*(4), 561–577.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Parabiaghi, A., Rossi, A. A., Castelnovo, A., del Fabro, L., Mannarini, S., Percudani, M. E., & for the CCM2013 Project team (corporate authorship) (2024). The Italian version of the 16-item Prodromal Questionnaire (PQ-16) and its psychometric features in help-seeking ultra-high-risk subjects and in the general population. *Early Intervention in Psychiatry*, 1–14. https://doi.org/10.1111/eip.13516