EDITORIAL

Screening instruments in child and adolescent psychiatry: general and methodological considerations

Chantal Michel · Frauke Schultze-Lutter ·
Benno G. Schimmelmann

Introduction

There is strong evidence that most psychiatric disorders have their origins early in life and that risk for psychiatric disorders in adulthood is increased by childhood adversities [1]. Furthermore, neuropsychiatric disorders are the most common causes of burden and disability in young persons aged 10–24 years in whom they account for 45 % of these, and are strongly associated with risk-behaviors and substantial psychosocial impairment [2–5]. An important concern is the duration of untreated illness which has been increasingly considered as a predictor of worse outcome across different psychiatric disorders [6]. Therefore, an early detection and adequate intervention are crucial to reduce overall burden and disability associated with neuropsychiatric disorders [7]. One important reason for the duration of untreated illness is that more than a third of patients with a psychiatric disorder do not or only with delay seek help from a mental health professional [8]. In contrast, most children and adolescents are regularly seen by general medical professionals for other reasons (e.g., primary care physician, pediatrician, or nurse) and/or by school counselors (pedagogues, social workers, or sometimes psychologists) if they have behavioral or emotional problems. These mainly non-mental health professionals need screening instruments to detect whether or not a child is in need for a general psychiatric evaluation (caseness) and, in the event that a specific psychiatric disorder is assumed, screeners for a particular disorder (e.g., ADHD,

C. Michel (✉) · F. Schultze-Lutter · B. G. Schimmelmann
Research Department, University Hospital of Child and
Adolescent Psychiatry and Psychotherapy, University of Bern,
Bolligenstrasse 111 (Haus A), 3000 Bern 60, Switzerland
e-mail: chantal.michel@kjp.unibe.ch

psychosis). Furthermore, even mental health professionals are in need for screens, if specialized, elaborate/sophisticated and/or time-consuming assessments are considered, e.g., for psychosis risk or autism [7, 9, 10].

Screenings are common in many areas of medicine, and screeners are frequently employed for the detection of psychiatric disorders [5, 9, 11]. However, in psychiatry, screening instruments are often discredited for their poor psychometric properties, such as too many false positives (i.e., poor positive predictive value) or lack of adaptions for certain age groups [9]. While some psychiatric disorders may indeed be difficult to screen for, the most serious problem is that reports on new screening instruments frequently lack sufficient evaluation of crucial psychometric properties that would be mandatory to judge their usefulness. This may have contributed to the bad reputation of psychiatric screening instruments.

Psychometric properties of screeners

Generally, data on reliability and validity as well as norms for the targeted population(s) is needed to evaluate its appropriateness.

Reliability relates to the accuracy of measurement by a screener—irrespective of whether or not it actually assesses the targeted construct. Three complementary aspects of reliability are distinguished: (1) *Test–retest* reliability requires that a screener should measure whatever it measures consistently over time (note: the test–retest reliability might well appear low when the screener measures a fluctuating state rather than a trait condition and when the condition itself has changed between test and retest assessment). (2) *Internal consistency* demands that items of the screener or its subscales are homogenous, i.e., measure

the same construct(s). (3) If the screen is an interview, *inter-rater reliability* evaluates the rate of agreement between different raters [12, 13].

Validity relates to the degree to that a screener actually measures what it is supposed to measure. Three complimentary main aspects of validity are commonly required for screening instruments: (1) of main interest for a clinical-diagnostic application is the *criterion validity*; it is an indicator of how well a screener's result corresponds to an individual result on a specified criterion [12, 13]. Thereby, two aspects of criterion validity are distinguished in relation to the time between the assessment of screener and criterion: (a) the degree to which the screener can identify individuals who currently have any or a specific psychiatric disorder (concurrent validity; requires nearly simultaneous assessment of screener and criterion in the test construction phase, while, later in practice, some time might pass between screening and formal psychiatric assessment), and (b) the extent to which an individual's score on a screener will accurately predict the individual's future result such as a psychiatric disorder (predictive validity; outcome criterion will reveal only in future and is assessed considerably later than screener) [12, 13]. (2) When the focus is less on the result of a screen but rather on its score and the measure of interest is less well defined than, for example, a formal diagnosis but relates to a construct that is not directly assessable (such as intelligence or personality characteristics), the *construct validity* is assessed. It refers to the extent to which screener scores correspond to scores of a gold standard assessment by expert consensus (such as the HAWIK in the assessment of IQ). One aspect of the construct validity is the convergent validity, which is good when the correlation between the screener and another established assessment of the same construct is high. The opposite aspect of construct validity is the discriminant validity, which is high when screener scores do not correlate with measures of other constructs [12]. For example, scores of an ADHD screener should not be highly positively correlated to scores of scales assessing oppositional defiant/conduct or emotional disorders. (3) The *content validity* finally requests that the screening instrument should measure all important aspects of the target condition, e.g., not only inattentiveness but also hyperactivity and impulsivity when ADHD and not only the inattentive subtype is targeted [12].

Taken together, besides producing (state) consistent scores and results (reliability), a screener must also be accurate with regard to content (validity). Generally reliability is more easily established than validity, and reliability is often first or exclusively described for instruments. Thus, many screeners lack validity data, which makes it difficult to know how clinically useful the instrument is [12, 13]. An ideal screening instrument for diagnostic purposes would demonstrate excellent concurrent (predictive) validity by (1) ruling in most if not all patients with the target condition (diagnosis) while (2) ruling out a considerable proportion of those without it. To rule in most patients with the target condition, a screener should generally possess a sensitivity approaching 100 %, a negative diagnostic likelihood ratio (LR) ≤0.1 that indicates a 'large and often conclusive' change from pre-screening to post-screening probability of the absence of illness risk [14], and a positive predictive value that is greatest in settings in which the prevalence of the condition is highest, i.e., greater in clinical settings than in community settings [15]. On the other hand, to rule out a considerable proportion of patients without the target condition, a screener should generally possess reasonably high specificity and a positive diagnostic LR ≥5 that indicates at least a moderate increase in the pre-screening to post-screening risk probability [14]. In many studies evaluating screening instruments, e.g., for psychosis risk [16], only sensitivity and specificity data are described, while diagnostic likelihood ratios are rarely provided, although these can be more easily interpreted [cutoffs for "good" concurrent (predictive) validity exist] and should, therefore, always be reported.

Additionally, the screener's differential accuracy should not be largely mediated by confounding conditions, e.g., comorbid emotional or behavioral disorders [17], but its items/components should possess good content, and convergent and/or criterion validity (i.e., indeed measure the target condition) [13]. For example, when tested alongside the gold standard of diagnosis in a clinical interview, the final screener result (e.g., determined by a cutoff score) should not only correspond to the interview result but also each single item of the screener should highly correlate with their respective interview counterparts (both are aspects of convergent validity in dimensional assessments or of criterion validity when presence of symptoms is rated) [13]. Further, all aspects and not only parts of the target condition should be assessed by the screening tool (content validity). These criteria are rarely addressed in studies evaluating screening instruments.

Last but not least for clinical purposes and the evaluation of the mental state of individual patients, norms or cutoffs should be provided that allow the evaluation of an individual performance against that of a similar group. To improve the population fit, screener should be adopted to the overall purpose (e.g., screening for psychiatric caseness in the general population vs. screening for a specific condition in a clinical population) or to different groups (e.g., separate norms for age groups, gender, and/or other potentially influential sociodemographic characteristics) [12].

Conclusions

Many studies on screening instruments lack the appropriate assessment of relevant psychometric properties. However, before studying psychometric properties, the purpose (e.g., caseness in general or a specific psychiatric diagnosis) and setting (e.g., general population/school, primary care or mental health services including the expected developmental stage of the recipients) of a screening should be clarified. Most screeners are not useful for all purposes (e.g., for caseness and a specific disorder). Consequently, psychometric properties should be studied in appropriate populations of adequate sample size using pre-defined (!) cutoffs for reliability and validity criteria (e.g., diagnostic likelihood ratios) that distinguish a useful from a useless screening instrument. Although it may be difficult to develop good screeners for all situations and conditions in child and adolescent psychiatry, and many studies of potential screeners are so far inappropriate, careful research on screening instruments is mandatory to improve comprehensive and early detection of psychiatric conditions in children and adolescents—in particular during times of increasingly tighter resources.

References

1. Kessler RC, Berglund P, Demler O, Jin R, Walters EE (2005) Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 62:593–602
2. Gore FM, Bloem PJN, Patton GC et al (2011) Global burden of disease in young people aged 10–24 years: a systematic analysis. Lancet 377:2093–2102
3. Kieling C, Baker-Henningham H, Belfer M et al (2011) Child and adolescent mental health worldwide: evidence for action. Lancet 378:1515–1525
4. Kaess M, Brunner R, Parzer P et al (2014) Risk-behaviour screening for identifying adolescents with mental health problems in Europe. Eur Child Adolesc Psychiatry 23(7):611–620
5. Polanczyk GV (2013) The burden of childhood mental disorders. Eur Child Adolesc Psychiatry 22(3):135–137
6. Dell'Osso B, Glick ID, Baldwin DS, Altamura AC (2013) Can long-term outcomes be improved by shortening the duration of untreated illness in psychiatric disorders? a conceptual framework. Psychopathology 46(1):14–21
7. Costello EJ, Egger H, Angold A (2005) 10-year research update review: the epidemiology of child and adolescent psychiatric disorders: I. Methods and public health burden. J Am Acad Child Adolesc Psychiatry 44(10):972–986
8. Alonso J, Angermeyer MC, Bernert S et al (2004) Use of mental health services in Europe: results from the European study of the epidemiology of mental disorders (ESEMeD) project. Acta Psychiatr Scand Suppl 420:47–54
9. Myers K, Winters NC (2002) Ten-year review of rating scales. I: overview of scale functioning, psychometric properties, and selection. J Am Acad Child Adolesc Psychiatry 41(2):114–122
10. Falkmer T, Anderson K, Falkmer M, Horlin C (2013) Diagnostic procedures in autism spectrum disorders: a systematic literature review. Eur Child Adolesc Psychiatry 22(6):329–340
11. Simon E, Dirksen CD, Bögels SM (2013) An explorative cost-effectiveness analysis of school-based screening for child anxiety using a decision analytic model. Eur Child Adolesc Psychiatry 22(10):619–630
12. Glover TA, Albers CA (2007) Considerations for evaluating universal screening assessments. J Sch Psychol 45:117–135
13. Dawe S, Loxton NJ, Hide L, Kavanagh DJ, Mattick RP (2002) Review of diagnostic screening instruments for alcohol and other drug use and other psychiatric disorders. Commonwealth of Australia, AU
14. Jaeschke R, Guyatt GH, Sackett DL (1994) Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? JAMA 271(9):703–770
15. Cadman D, Chambers L, Feldman W, Sackett D (1984) Assessing the effectiveness of community screening programs. JAMA 251(12):1580–1585
16. Kline E, Schiffman J (2014) Psychosis risk screening: a systematic review. Schizophr Res. doi:10.1016/j.schres.2014.06.036
17. Janes H, Pepe MS (2008) Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting. Am J Epidemiol 16(1):89–97