Psychological Medicine

cambridge.org/psm

The development and validation of an algorithm to predict future depression onset in unselected youth

Joseph R. Cohen[1], Hena Thakur[1], Jami F. Young[2] and Benjamin L. Hankin[1]

[1]Department of Psychology, University of Illinois Urbana-Champaign, Champaign, IL USA and [2]Department of Child and Adolescent Psychiatry and Behavioral Sciences, Children's Hospital of Philadelphia, Philadelphia, PA USA

Original Article

Cite this article: Cohen JR, Thakur H, Young JF, Hankin BL (2020). The development and validation of an algorithm to predict future depression onset in unselected youth. *Psychological Medicine* **50**, 2548–2556. https://doi.org/10.1017/S0033291719002691

Received: 20 May 2019
Revised: 11 July 2019
Accepted: 8 September 2019
First published online: 2 October 2019

Key words:
Assessment; pediatric depression; receiver operating characteristics; screening

Author for correspondence:
Joseph R. Cohen, E-mail: cohenj@illinois.edu

Abstract

Background. Universal depression screening in youth typically focuses on strategies for identifying current distress and impairment. However, these protocols also play a critical role in primary prevention initiatives that depend on correctly estimating future depression risk. Thus, the present study aimed to identify the best screening approach for predicting depression onset in youth.

Methods. Two multi-wave longitudinal studies ($N = 591$, $Age_M = 11.74$; $N = 348$, $Age_M = 12.56$) were used as the 'test' and 'validation' datasets among youth who did not present with a history of clinical depression. Youth and caregivers completed inventories for depressive symptoms, adversity exposure (including maternal depression), social/academic impairment, cognitive vulnerabilities (rumination, dysfunctional attitudes, and negative cognitive style), and emotional predispositions (negative and positive affect) at baseline. Subsequently, multi-informant diagnostic interviews were completed every 6 months for 2 years.

Results. Self-reported rumination, social/academic impairment, and negative affect best predicted first depression onsets in youth across both samples. Self- and parent-reported depressive symptoms did not consistently predict depression onset after controlling for other predictors. Youth with high scores on the three inventories were approximately twice as likely to experience a future first depressive episode compared to the sample average. Results suggested that one's likelihood of developing depression could be estimated based on subthreshold and threshold risk scores.

Conclusions. Most pediatric depression screening protocols assess current manifestations of depressive symptoms. Screening for prospective first onsets of depressive episodes can be better accomplished via an algorithm incorporating rumination, negative affect, and impairment.

Depression is a significant pediatric public health concern in the USA as over three million adolescents experience a depressive episode annually (Weinberger *et al.*, 2017). Yet, millions of depressed youth go unrecognized by health providers, and consequentially, remain untreated for this chronic, impairing, and potentially fatal mental illness (Kessler *et al.*, 2001). In response, several governmental [e.g. the United States Preventive Services Task Force (USPSTF); Siu, 2016] and professional (e.g. the American Academy of Pediatrics; Zuckerbrot *et al.*, 2018) organizations now recommend universal depression screening starting at age 12. According to the USPSTF, universal depression screening has two aims: (1) identify current distress and impairment, and (2) estimate prospective depression risk (Siu, 2016). To date, most research focuses on how well depressive symptom measures classify current states of clinical depression, with few studies examining if these same protocols predict future depression outcomes (see Stockings *et al.*, 2015). Thus, translational research demonstrating the clinical utility of methods for predicting future incidents of depression is needed.

Developing a screening protocol for future depression onset (i.e. a first lifetime episode of depression) may be particularly important. Depression is a recurrent disorder, and prevention efforts demonstrate better long-term outcomes compared to depression interventions (Weersing *et al.*, 2017). Primary prevention screening can therefore help reduce the burden of depression by preventing its onset. Conceptual models (e.g. the kindling hypothesis) posit that risk factors for first lifetime and recurrent depression episodes may vary (Monroe and Harkness, 2011). However, the relatively infrequent use of diagnostic assessments in large, multi-wave studies makes it challenging to disentangle which risk factors are best to preselect for prospectively predicting first depression onset in adolescents.

In addition to differentiating between recurrent and first lifetime episodes of depression, additional translational barriers exist. First, few studies simultaneously examine risk factors, making it challenging to perform formal tests of incremental validity (Johnston and Murray, 2003). Second, most studies rely on prediction models that use regression-based, structural equation modeling, or machine learning analytic plans. While all of these broad analytic frameworks have significant, and unique, strengths, they often do not speak to all aspects

CAMBRIDGE UNIVERSITY PRESS

of predictive validity when used in isolation. Specifically, (a) predictive accuracy (e.g. the ability to differentiate between those who will and will not develop depression; also known as *discrimination*) and (b) precise likelihood estimates (e.g. correctly estimating the risk of depression onset for a range of scores; also known as *calibration*) should be explicitly tested to demonstrate the clinical usefulness of a prediction model (Hanson, 2016; Youngstrom *et al*., 2017; Lindhiem *et al*., 2018). Finally, the lack of replication studies is a critical issue for psychological research, including mental health screening. Without replication, prediction models will likely (unintentionally) exaggerate a given screening protocol's ability to identify the risk (Steyerberg, 2009).

Beyond statistical prediction, there are other factors to consider when selecting screening measures. For instance, early childhood adversities predict first lifetime episodes (Monroe and Harkness, 2011), but focusing on historical events could be stigmatizing and does not help explain the dynamic reason *why* certain adolescents may be at increased risk (Finkelhor, 2018). Instead, psychosocial individual differences, such as cognitive vulnerabilities (e.g. rumination) and emotional predispositions (e.g. negative affect), may be preferable risk factors to target in screening batteries because they can be altered by evidence-based therapeutic approaches (Garber *et al*., 2012; Nehmy and Wade, 2015). Similarly, academic and social difficulties, two forms of impairment associated with adolescent depression (Jaycox *et al*., 2009), can also be attenuated via improved problem-solving skills often taught in depression prevention protocols. By focusing on dynamic personal factors that are targeted in existing prevention interventions, the screening process can select youth based on these proposed mechanisms of risk.

In the present study, we sought to develop and validate an algorithm for prospectively predicting first depression onset in youth. To do so, we analyzed data from two independent, multiwave longitudinal studies with shared methodological characteristics and multiple psychosocial risk factors (e.g. cognitive vulnerabilities, emotional predispositions, impairment, maternal depression). The primary aims for this research are to (a) bolster pediatric depression screening initiatives and (b) advance clinical research. Many randomized clinical trials for depression prevention programs use subthreshold depressive symptoms to select vulnerable youth (Weersing *et al*., 2017). The current study is among the first to comprehensively test whether this is the best approach. Moreover, our study is well-positioned to quantify the incremental validity and clinical utility of using multiple predictors to forecast depression onset (i.e. a multi-indicator screening approach). Across screening initiatives for pediatric conditions (e.g. obesity; Proctor *et al*., 2003), it is well-accepted that multiple indices are necessary for predicting chronic health outcomes. Screening batteries that quantify the risk across several predictors provide flexibility in the decision model and help protect against the error associated with depending on a single cutoff (Cohen *et al*., 2018*a*). Thus, we seek to identify a screening algorithm to predict future first depression onset that is both accurate and clinically useful.

Method

Participants and procedure

The Gene-Environment and Mood (GEM) and Montreal-Chicago (MTL-CHI) studies, two independent, multi-site samples were used to develop our algorithms. For both studies, youth were recruited by partnering with schools and sending letters home to parents, and via other indirect methods (e.g. advertisements placed in local newspapers). Youth were excluded if the parent reported that the youth (a) had an intellectual disability that prevented them from understanding the questionnaires, (b) suffered from psychosis, or (c) did not speak English. For the GEM study, 663 parent–child dyads enrolled, while the MTL-CHI study consisted of 382 dyads. At baseline, youth and a caregiver completed self- and parent-reported depressive symptoms, adversities, maternal depression, impairment, and individual differences (cognitive vulnerabilities, emotional predisposition) inventories. In addition, four diagnostic assessments in a 2-year span (i.e. 6-month follow-up interviews) were completed to assess for depression onset. The GEM study took place from 2008 to 2013 and the MTL-CHI took place from 2003 to 2007.

Only youth who, as determined via our multi-informant diagnostic interview, had not experienced a major depressive episode (MDE) in their lifetime were used in the current study. This resulted in a subsample of 591 youth in GEM and 348 in MTL-CHI. The distribution of age (in years; GEM: $M = 11.74$, range = 7–17; MTL-CHI: $M = 12.56$, range = 10–15), sex (GEM: 55.2% female; MTL-CHI: 56.6% female), and race (GEM: White = 62.2%, African-American = 12.3%; MTL-CHI: White = 70.6%, African-American = 12.3%) was similar between the studies (see Hankin *et al*. (2015) and Abela and Hankin (2011) for more details on the GEM and MTL-CHI studies). Identical measures were used to assess predictors and depression onset in both studies.

Measures

Depression diagnoses

Trained clinical psychology doctoral students administered the Mood Disorders section of the Schedule for Affective Disorders and Schizophrenia for School Age Children (K-SADS-PL) (Kaufman *et al*., 1997) to youth and a caretaker at baseline and each follow-up. Interviewers were trained and supervised by licensed clinical psychologists. Both interviews were used to determine youths' diagnostic status using best estimate procedures (Klein *et al*., 2005). Interviewers were blinded to all other study procedures when making a diagnosis. Inter-rater reliability for the KSADs was excellent for both GEM ($\kappa = 0.91$) and MTL-CHI ($\kappa = 0.87$) based on approximately 20% of interviews reviewed for reliability. Youth were diagnosed with depression if they met DSM criteria for at least five symptoms of an MDE. In total, 5.25% ($N = 31$) and 8.30% ($N = 29$) of youth experienced depression onset in the GEM and MTL-CHI studies, respectively. This rate of depression onset in a 2-year period is comparable to past research in a similar age range (Hankin *et al*., 1998). The KSADs is a common diagnostic assessment of youth psychiatric disorders and has demonstrated excellent reliability and validity for depression outcomes (Hankin and Cohen, in press).

Depression symptoms

The Children's Depression Inventory (CDI) (Kovacs, 1992) assessed both self- and parent-reported symptoms in youth. The youth (CDI-Y) and parent (CDI-P) report on the CDI are identical except parents answer with regard to how they believe their child feels. Scores on the CDI-Y and CDI-P are reliable and valid predictors of depression in youth (Garber, 1984; Kovacs, 1992). The CDI is a commonly used screening inventory for depression, with similar predictive validity as other depression

https://doi.org/10.1017/S0033291719002691 Published online by Cambridge University Press

inventories (Hankin and Cohen, in press). The CDI-Y ($\alpha = 0.88$ in GEM; $\alpha = 0.87$ in MTL-CHI) and CDI-P ($\alpha = 0.86$ in GEM; $\alpha = 0.85$ in MTL-CHI) had acceptable levels of reliability in the current study.

Adversities

Items from the Adolescent Life Event Questionnaire (ALEQ) (Hankin and Abramson, 2002) were used to assess recent adversity exposure. Proximal adversity exposure (past 3 months) was intentionally queried because recent adversity exposure may be uniquely predictive for depression onset (Monroe and Harkness, 2011). The subscale asked about the youth's direct and/or indirect exposure to a number of potentially traumatic events (i.e. any hospitalizations, deaths, arrests, significant medical or emotional problems, parental separation/divorce, job loss, and emotional neglect exposure).

Parental depression

One risk factor that may be particularly important to assess for depression screening in youth is parental depression (Siu, 2016). In response, the Beck Depression Inventory (Beck *et al.*, *1996*) was used to assess current depressive symptoms and the Schedule of Clinical Diagnoses (SCID) (Lobbestael *et al.*, 2011) was used to assess lifetime history of parental depression. Both the BDI and SCID are valid and reliable measures of adult depression (Pettersson *et al.*, 2015). Overall, 27% ($N = 160$) and 32% ($N = 112$) of participating caretakers reported a history of major depression in the GEM and MTL-CHI studies, respectively. BDI and SCID scores were used as independent indices of parental depression.

Social and academic impairment

Another ALEQ (Hankin and Abramson, 2002) subscale measured academic and social impairment. Examples of academic impairment included receiving a bad report card and having a bad teacher/class. Social impairment reflected social isolation (e.g. not having as many friends as you'd like to) and conflict (e.g. arguments with parents) with peers and family members. Impairment was conceptualized as a multi-dimensional measure that assessed both social and academic impairment (rather than social *or* academic impairment) based on recommendations within the field (Fabiano and Pelham, 2016). The ALEQ is a reliable and valid measure of items related to adolescent impairment (Hankin *et al.*, 2010).

Individual differences

Several psychosocial risk factors were assessed via self-report questionnaires. These include the Children's Dysfunctional Attitudes Scale (Abela and Sullivan, 2003), the Adolescent Cognitive Style Questionnaire (Hankin and Abramson, 2002) (to assess inferential styles), and the Children's Response Style Questionnaire (Abela *et al.*, 2004) (to assess rumination). These inventories collectively measure cognitive styles articulated in preeminent theoretical models that explain depression onset in adolescents (Hankin *et al.*, 2016). Additionally, the Positive and Negative Affect Scale for Children (Laurent *et al.*, 1999) was included as a measure of positive and negative affect, both of which are temperamental risk factors for depression onset in youth (Muris and Ollendick, 2005; Farchione *et al.*, 2012). In the current study, Cronbach α for dysfunctional attitudes ($\alpha = 0.85$ in GEM; $\alpha = 0.67$ in MTL-CHI), inferential style ($\alpha = 0.93$ in GEM; $\alpha = 0.90$ in MTL-CHI), rumination ($\alpha = 0.87$ in GEM; $\alpha = 0.89$ in MTL-CHI), positive affect ($\alpha = 0.83$ in GEM; $\alpha = 0.90$ in MTL-CHI), and negative affect ($\alpha = 0.89$ in GEM; $\alpha = 0.88$ in MTL-CHI) were acceptable and consistent with past research (Laurent *et al.*, 1999; Muris and Ollendick, 2005; Abela and Scheffler, 2008; Hankin *et al.*, 2016).

Data analytic plan

As the GEM study had a larger sample size, it represented our 'test' study, while MTL-CHI was our 'validation' study. Analyses first examined the properties of *each* measure dimensionally to establish the predictive validity of a given index test (Youngstrom *et al.*, 2017). Specifically, because few translational studies have used a risk factor approach to depression screening, our initial aim was to test, and subsequently validate, each index test (i.e. a *pre-selection* approach; Steyerberg, 2009). Once a collective set of valid predictors was identified, we subsequently conducted multivariate analyses to test and validate if the multi-indicator screening algorithm was incrementally valid across both datasets. To minimize potential bias in our parameter estimates, bias-corrected accelerated bootstrap methods were used when appropriate.

There are a variety of analytic approaches that can quantify the utility of a prediction model. In the extant literature, various machine learning and related data mining approaches have become common exploratory steps in identifying a subset of novel predictors. However, given that we were focused on the clinical utility of a finite number of relatively well-known risk factors, the advantages of a machine learning approach are mitigated in the current context. Instead, we began with a regression-based approach as it has shown to be more robust compared to alternative methods (e.g. machine learning) within studies that have comparable methodological characteristics (i.e. <25 predictors and sample sizes 1000) (Steyerberg, 2009). Specifically, discrete-time survival models (Singer and Willett, 2003) were first conducted to examine which measures predicted time to first episode onset.

Next, we used a 'best practice' receiver operator characteristic (ROC) approach (Youngstrom, 2014) to compute the AUC for each measure. Across both machine learning and regression-based approaches, ROCs are useful in determining whether a predictor is not only statistically significant, but also clinically useful (Youngstrom *et al.*, 2017). If a given predictor was significant in both the survival analyses and ROC, it was deemed to have satisfactory predictive accuracy. Subsequently, Diagnostic Likelihood Ratios (DLRs) were computed to estimate the likelihood of developing depression based on minimal, subthreshold, and threshold levels of risk. Consistent with an evidence-based medicine approach (Straus *et al.*, 2011), these groups were formed by using cutoffs that generated three equal groups of adolescents. Finally, the Expected-Observed (E/O) Index was then used to determine whether the likelihood estimates generated by the DLRs were valid (Hanson, 2016). Specifically, the DLRs for each risk profile (e.g. threshold rumination) in the GEM study were multiplied by the pre-test odds for depression onset in the MTL-CHI study. This represented the *expected* number of episodes given the number of youth with that risk profile in the MTL-CHI study. We then divided this number by the *observed* number of episodes for that profile in the MTL-CHI study. The Poisson variance for the logarithm of the observed number of cases was used to create the confidence interval (CI) for the E/O Index. A CI that included 1 indicates strong calibration (Hanson, 2016).

Several tests of incremental validity were next conducted with our valid predictors. First, discrete-time survival analyses using

https://doi.org/10.1017/S0033291719002691 Published online by Cambridge University Press

Table 1. Descriptive statistics for multi-site study

		GEM study		MTL-CHI study	
	# of items	Mean	S.D.	Mean	S.D.
CDI-Y	27	6.51	(5.21)	8.81	(6.53)
CDI-P	27	4.18	(4.57)	6.24	(5.58)
Adversities	6	11.22	(3.77)	11.97	(4.27)
BDI	21	5.33	(6.32)	6.89	(7.43)
Impairment	18	32.19	(9.00)	36.12	(10.22)
Dysfunctional attitudes	9	33.46	(7.42)	32.11	(5.90)
Attributional style	9	2.87	(0.87)	2.51	(0.78)
Rumination	25	25.74	(7.30)	26.21	(7.23)
Positive affect	12	44.72	(8.38)	44.45	(8.29)
Negative affect	15	27.21	(9.22)	28.54	(9.46)

CDI-Y, Children's Depression Inventory (Kovacs, 1992) (Youth Report); CDI-P, Children's Depression Inventory (Kovacs, 1992) (Parent Report); Adversities, Adversity subscale of the Adolescent Life Event Questionnaire (Hankin and Abramson, 2002); BDI, Beck Depression Inventory (Beck *et al*., 1996); Impairment, Impairment subscale of the Adolescent Life Event Questionnaire (Hankin and Abramson, 2002); Dysfunctional Attitudes, Children's Dysfunctional Attitudes Scale (Abela and Sullivan, 2003); Attributional style, Attributional Cognitive Style Questionnaire (Hankin and Abramson, 2002); Rumination, Children's Rumination Scale Questionnaire (Abela *et al*., 2004); Positive affect, Positive and Negative Affect Scale for Children (Laurent *et al*., 1999) (Positive Affect Subscale); Negative affect, PANAS (Laurent *et al*., 1999) (Negative Affect Subscale).

backward selection methods based on changes in the maximum likelihood estimates of the likelihood ratio were examined (Steyerberg, 2009). If a predictor was eliminated across both studies, it was dropped from further analyses. Further, AUCs of remaining predictors were compared using the DeLong test for paired ROC curves (Youngstrom *et al*., 2017). A significant difference in the DeLong test suggests that certain index tests should be prioritized in prediction tools.

Finally, an additive approach was used to form a Cumulative Risk score. Specifically, profiles were formed based on whether youth were at minimal risk (0), subthreshold risk (1), or threshold risk (2) on each measure. This categorical approach is similar to how cumulative risk is typically conceptualized within basic research (e.g. Evans, 2003) and in multi-indicator risk algorithms for pediatric health screening (e.g. Proctor *et al*., 2003). We then replicated the analytic plan described above to test whether (a) the Cumulative Risk score (comprised of the categorical risk score on each index test) predicted depression onset above and beyond the individual index tests and (b) the Cumulative Risk score prospectively predicted depression onset above and beyond current approaches that rely on depressive symptoms. Additional calibration plots were created to test whether the predicted probabilities for the Cumulative Risk score mirrored the observed probabilities within each dataset (Lindhiem *et al*., 2018). Online Supplementary Table S1 summarizes our analytic approach. The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines were followed in reporting this study's methods and results (Collins *et al*., 2015).

Results

Missing data and preliminary analyses

As is common in longitudinal studies, there were missing data across the follow-ups. Little's Missing Completely at Random (MCAR) test suggested that these data were missing completely at random in both the GEM [$\chi^2(35) = 21.76$, $p = 0.96$] and MTL-CHI [χ^2 (173) = 201.10, $p = 0.07$] studies. Thus, missing baseline data were imputed using expectation maximization algorithms for the predictor, and follow-up data were censored after the last completed follow-up. Descriptive statistics for our dimensional predictors can be found in Table 1. Online Supplementary Fig. S1 displays Kaplan–Meier curves to illustrate the rate of depression onset in our study.

Univariate analyses

Results from univariate survival analyses and ROC are presented in Table 2. Based on the Wald statistic and AUC, only three significant predictors from the GEM study were also significant in the MTL-CHI study: Rumination, Impairment, and Negative Affect. Of note, both self- and parent-reported depressive symptoms demonstrated inconsistent findings across the two studies. DLRs and the accompanying E/O indices for Rumination, Impairment, and Negative Affect are presented in Table 3, along with the corresponding cutoffs for minimal, subthreshold, and threshold levels of risk. Importantly, youth were approximately 1.5–2 times more likely than the sample average to be diagnosed with a first lifetime episode of depression if they scored in the upper third on any one of the three measures. CIs of the E/O Index suggest that each predictor was able to reliably estimate the likelihood of depression onset for minimal, subthreshold, and threshold risk levels.

Multivariate analyses

When entered simultaneously into multivariate survival analytic models, neither Rumination, Impairment, nor Negative Affect consistently emerged as incrementally significant predictors using backward selection methods. Further, DeLong tests indicated that there were no significant differences between the AUCs for our predictors ($p > 0.05$). We therefore included, and equally weighted, Rumination, Impairment, and Negative Affect to form the Cumulative Risk score. This created a range of scores from 0 (minimal risk across each predictor) to 6 (elevated risk for

https://doi.org/10.1017/S0033291719002691 Published online by Cambridge University Press

Table 2. Univariate discrete-time survival analyses and receiver operating characteristics

	GEM study				MTL-CHI			
	Wald	*p*	OR (CI 95%)	AUC (CI 95%)	Wald	*P*	OR (CI 95%)	AUC (CI 95%)
CDI-Y	11.77	0.001	1.10 (1.04–1.17)	0.67** (0.58–0.77)	1.89	0.17	1.04 (0.99–1.09)	0.64** (0.56–0.72)
CDI-P	0.22	0.64	1.00 (0.99–1.00)	0.72** (0.62–0.82)	1.42	0.23	1.04 (0.98–1.10)	0.63** (0.55–0.71)
Adversities	2.43	0.12	1.06 (0.99–1.14)	0.59 (0.50–0.69)	1.24	0.27	1.04 (0.97–1.13)	0.62** (0.54–0.71)
BDI	6.65	0.01	1.06 (1.01–1.10)	0.64** (0.54–0.73)	0.010	0.93	1.00 (0.96–1.05)	0.53 (0.44–0.62)
Dysfunctional attitudes	0.06	0.81	1.00 (0.98–1.02)	0.53 (0.45–0.62)	0.52	0.47	1.02 (0.96–1.09)	0.54 (0.45–0.63)
Attributional style	0.05	0.83	1.00 (0.98–1.02)	0.56 (0.45–0.67)	2.31	0.13	1.41 (0.91–2.18)	0.55 (0.45–0.65)
Rumination	4.61	0.03	1.05 (1.00–1.10)	0.64** (0.56–0.72)	8.04	0.005	1.07 (1.02–1.12)	0.68** (0.60–0.77)
Positive affect	1.16	0.28	0.98 (0.94–1.02)	0.53 (0.35–0.58)	0.32	0.57	0.99 (0.94–1.03)	0.44 (0.34–0.54)
Negative affect	9.92	0.002	1.05 (1.02–1.08)	0.67** (0.58–0.77)	4.24	0.04	1.04 (1.00–1.08)	0.65** (0.57–0.73)
Impairment	14.75	<0.001	1.07 (1.03–1.11)	0.70** (0.61–0.77)	4.71	0.03	1.04 (1.00–1.08)	0.68** (0.60–0.75)
Parent depression Dx	0.12	0.73	1.00 (0.98–1.02)	0.60 (0.49–0.71)	0.003	0.96	1.02 (0.47–2.21)	0.52 (0.43–0.62)

CDI-Y, Children's Depression Inventory (Kovacs, 1992) (Youth Report); CDI-P, Children's Depression Inventory (Kovacs, 1992) (Parent Report); subscale Adversities, Adversities subscale of the Adolescent Life Event Questionnaire (Hankin and Abramson, 2002); BDI, Beck Depression Inventory (Beck *et al*., 1996); II Dysfunctional attitudes, Children's Dysfunctional Attitudes Scale[31]; Attributional style, Attributional Cognitive Style Questionnaire (Hankin and Abramson, 2002); Rumination, Children's Rumination Scale Questionnaire (Abela *et al*., 2004); Positive affect, Positive and Negative Affect Scale for Children (Laurent *et al*., 1999) (1999) – Positive Affect Subscale; Negative affect, PANAS (Laurent *et al*., 1999) – Negative Affect Subscale. Impairment, Impairment subscale of the Adolescent Life Event Questionnaire (Hankin and Abramson, 2002); Parent depression Dx, Lifetime depression diagnosis from the depression module of the SCID. All estimates from both the survival and receiver operating characteristic analyses are bootstrapped.
**$p \leqslant 0.01$.

Table 3. Diagnostic Likelihood Ratio and expected/observed indices

Predictor (risk level)	Score range	DLR (GEM) (95% CI)	DLR (MTL-CHI) (95% CI)	E/O Index (95% CI)
Rumination (minimal)	⩽21	0.28 (0.10–0.84)	0.43 (0.17–1.09)	1.00 (0.44–2.27)
Rumination (average)	22–27	1.11 (0.65–1.88)	0.81 (0.43–1.48)	1.25 (0.70–2.23)
Rumination (elevated)	⩾28	1.58 (1.15–2.17)	1.73 (1.23–2.44)	0.88 (0.59–1.31)
Negative affect (minimal)	⩽21	0.38 (0.15–0.97)	0.49 (0.19–1.23)	1.00 (0.33–1.70)
Negative affect (average)	22–29	0.67 (0.62–0.70)	0.96 (0.57–1.61)	1.22 (0.48–1.34)
Negative affect (elevated)	⩾30	1.96 (1.47–2.62)	1.45 (0.99–2.12)	1.32 (0.83–1.94)
Impairment (minimal)	⩽26	0.10 (0.01–0.71)	0.15 (0.02–1.07)	1.00 (0.19–5.16)
Impairment (average)	27–35	1.15 (0.75–1.77)	0.87 (0.41–1.84)	1.22 (0.71–2.11)
Impairment (elevated)	⩾36	1.71 (1.21–2.40)	1.51 (1.13–2.03)	1.32 (0.91–1.92)

Rumination, Children's Rumination Scale Questionnaire (Abela *et al*., 2004); Negative affect, PANAS – Negative Affect Subscale (Laurent *et al*., 1999); Impairment, Impairment subscale of the Adolescent Life Event Questionnaire (Hankin and Abramson, 2002); DLR, Diagnostic Likelihood Ratios (Straus *et al*., 2011); E/O Index, Expected number of cases in the MTL-CHI study based on posterior probabilities for a given risk profile derived from the GEM study/the number of actual observed cases for a given risk profile in the MTL-CHI study. 95% Confidence Interval, The confidence interval for the E/O Index (if the CI includes 1, it is significant) (Hanson, 2016).

each predictor) for each youth-based one's level of risk (0–2) on the three predictors.

We next tested the incremental validity of our Cumulative Risk score (see Table 4). First, Cumulative Risk was entered into a model with Rumination, Negative Affect, and Impairment as individual predictors, as well as the CDI-Y and CDI-P, two traditional approaches to depression screening (Siu, 2016). Cumulative Risk was selected as the only significant prospective predictor of time until first depression onset in the GEM study, and these findings were validated in the MTL-CHI study. Next, to provide a more rigorous test of incremental validity, we created a Multi-Informant Risk score (summing minimal, subthreshold, and threshold risk across the CDI-Y and CDI-P). The purpose of these analyses was to examine our Cumulative Risk score against a 'gold standard' multi-informant approach (Klein *et al*., 2005). As there were only two predictors entered, we examined boostrapped estimates of the Cumulative Risk and Multi-Informant scores simultaneously. Findings suggested that both predictors were unique in predicting time to first onset in the GEM study, but only Cumulative Risk predicted time to first onset in the MTL-CHI study. AUCs for Cumulative Risk were 0.73 and 0.68, respectively, the highest AUCs of any predictor for each study.

Finally, we examined the E/O Index for our Cumulative Risk score. As can be seen in online Supplementary Fig. S2, the estimated likelihood matched the observed likelihood across the range of scores. Relatedly, the predicted probabilities were similar to the observed probabilities in both studies (online Supplementary Fig. S3). These findings suggest the Cumulative

https://doi.org/10.1017/S0033291719002691 Published online by Cambridge University Press

Table 4. Incremental validity and screening properties of risk score algorithm

GEM study				MTL-CHI study			
Predictor	Wald	*p*	OR (95%)	Predictor	Wald	*P*	OR (95%)
Cumulative risk score (0–6) compared to dimensional individual predictors of depression onset							
Impair	<0.01	0.96	1.00 (0.94–1.06)	Negative affect	0.10	0.93	1.00 (0.94–1.06)
CDI-P	0.26	0.61	1.00 (1.00–1.00)	Impairment	0.02	0.88	1.00 (0.94–1.05)
Negative affect	0.42	0.52	0.98 (0.93–1.04)	Rumination	0.34	0.56	1.02 (0.95–1.10)
CDI-Y	2.01	0.16	1.05 (0.98–1.12)	CDI-P	0.58	0.45	1.03 (0.96–1.09)
Rum	1.60	0.21	0.95 (0.88–1.03)	CDI-Y	0.53	0.47	0.98 (0.91–1.04)
Risk score	17.21	<0.001	1.64 (1.30–2.07)	Risk score	10.39	0.001	1.47 (1.16–1.85)
Cumulative risk score (0–6) compared to multi-informant score (0–4) for predicting depression onset							
Risk score	7.02	0.008	1.40 (1.09–1.80)	Risk score	7.15	0.008	1.42 (1.10–1.83)
Multi-informant score	8.38	0.004	1.67 (1.18–2.37)	Multi-informant score	0.50	0.48	1.13 (0.79–1.60)

Screening metrics

Risk level (score)	DLR (Comb.)	E/O Index	Sensitivity (GEM|M-C)		Specificity (GEM|M-C)		PPV (GEM|M-C)		NPV (GEM|M-C)	
Minimal (0–2)	0.30 (0.15–0.59)	1.00 (0.32–3.10)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Subthreshold (3–4)	0.91 (0.61–1.35)	1.00 (0.52–1.92)	0.87 (0.75–0.99)	0.90 (0.79–0.99)	0.42 (0.38–0.46)	0.35 (0.30–0.41)	0.08 (0.05–0.10)	0.11 (0.07–0.15)	0.98 (0.97–0.99)	0.97 (0.95–0.99)
Threshold (5–6)	2.13 (1.67–2.70)	1.18 (0.73–1.89)	0.58 (0.41–0.75)	0.59 (0.41–0.77)	0.75 (0.71–0.78)	0.69 (0.64–0.74)	0.11 (0.06–0.16)	0.15 (0.08–0.21)	0.97 (0.95–0.99)	0.95 (0.92–0.98)

Parameters for discrete-time survival analyses (DTSA) are derived from two separate models. In the top panel, parameters represent estimates from successive backward elimination models. The middle panel represents parameter estimates from simultaneous DTSA models in the GEM and MTL-CHI studies, respectively. In the top two panels, Risk score represents the sum of whether one was at minimal (0), subthreshold (1), or threshold (2) levels of risk. In the bottom panel, Risk score represents minimal (dimensional scores 0–2), subthreshold (3–4), and threshold levels of risk. The equation for the Diagnostic Likelihood Ratio (DLR) is: (number of episodes in risk profile/number of episodes total)/(number of non-episodes in risk profile/number of non-episodes total) (Straus *et al.*, 2011). DLRs above are combined between the two studies and weighted by the number of individuals in each study. Base rates for depression onset in GEM and MTL-CHI were 5.25% and 8.30%, respectively.

Risk score can be used as a dimensional prospective predictor of first depression onset (Hanson, 2016). Alternatively, it may be useful to have cutoffs that correspond to minimal, subthreshold, and threshold risk to further facilitate clinical decision-making (Straus *et al.*, 2011). The bottom panel of Table 4 provides the DLRs and E/O indices, as well as other complementary statistics commonly used in the screening literature [sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV)], for those at minimal risk, subthreshold risk, and threshold risk. Online Supplementary Fig. S3 shows how many individuals presented at each risk level across both studies. E/O indices at the bottom of Table 4 suggests the categorical version of the algorithm is well-calibrated. Online Supplementary Figure 4 provides the survival curves based on these cutoffs. This figure illustrates that threshold scores on the Cumulative Risk score developed in the GEM study were at an increased risk for depression onset in the MTL-CHI study.

Discussion

Increasingly, agencies and organizations are recommending pediatric depression screening with the aim of identifying current distress and impairment as well as future risk (Siu, 2016; Zuckerbrot *et al.*, 2018). However, few studies examine the optimal method for screening for prospective first onsets of depression. Overall, our findings suggest that relying on depression symptom inventories may lead to inconsistent findings when predicting future depression onset. Instead, measures of rumination, impairment, and negative affect represent reliable, clinically useful options for prospectively predicting depression onsets in youth. Below, we discuss the implications of this algorithm, and explain how these findings can help bridge the translational gap by providing prevention initiatives with a feasible solution for quantifying depression risk.

A key emphasis within the assessment literature is developing algorithms that accurately predict risk with results that replicate across independent samples (Steyerberg, 2009). Whereas youth and parent-reported depressive symptoms predicted depression onset in the test sample, these results did not replicate in the validation sample. Meanwhile, rumination, negative affect, and impairment all emerged as significant predictors that were incrementally valid compared to multi-informant symptom approaches. Our findings are consistent with theoretical models and corresponding basic research that conceptualizes these three risk factors as preceding depression onset (Laurent *et al.*, 1999; Muris and Ollendick, 2005; Abela and Scheffler, 2008; Hankin *et al.*, 2016). Yet, methodological limitations have inhibited the clinical utility of these measures, despite recent calls to integrate a focus on risk factors into depression prevention efforts (Garber *et al.*, 2012; Siu, 2016). By examining these risk factors within the context of other viable predictors of depression onset (e.g. maternal depression), and using a translational analytic

https://doi.org/10.1017/S0033291719002691 Published online by Cambridge University Press

Table 5. Diagnostic Likelihood Ratios and posterior probabilities for example screening cases

Examples of screening cases				
	Pre-test probability	Scoring profile	DLR	Overall post-test probability
Girl A (14 y/o)	8.33%	Rumination: 2 Negative affect: 2 Impairment: 1 Total risk score: 5	1.69	13.34%
Girl B (14 y/o)	8.33%	Rumination: 1 Negative affect: 2 Impairment: 2 Total risk score: 5	1.69	13.34%
Girl C (14 y/o)	8.33%	Rumination: 2 Negative affect: 2 Impairment: 2 Total risk score: 6	2.49	18.46%

Rumination, Children's Response Style Questionnaire (Abela *et al*., 2004); Negative affect, Positive and Negative Affect Schedule (PANAS) for Children Score (Laurent *et al*., 1999); Impairment, 18-item subscale from the Adolescent Life Event Questionnaire (ALEQ); DLR, Diagnostic Likelihood Ratio [(number of episodes in risk profile/number of episodes total)/(number of non-episodes in risk profile/number of non-episodes total)] (Straus *et al*., 2011); Pre-test probability, The baseline percentage of 14 year-old females in the Gene-Environment Mood study with first lifetime episodes within a year period; Post-test probability [(prevalence/(1–prevalence) × DLR)/((prevalence/(1–prevalence)) + 1] (Straus *et al*., 2011). Scores following our individual predictors represent the score assigned based on whether their score fell in the average (1) range or elevated (2) range. The Total risk score is derived by summing the scores across these three predictors. DLRs are the combined estimates from the GEM and MTL-CHI studies (weighted by the number of youth in each study). Although the development of the algorithm did not take into account gender and age, clinicians are able to account for these differences through use of an informed pre-test probability (as done above) (Youngstrom *et al*., 2017). Scores for Girl C illustrates the importance of undertaking an actuarial approach, in which, compared to Girls A and B, a Total risk score difference of 1 can dictate more intensive follow-up mental health services. Girls A and B, on the other hand, demonstrate an instance in which scoring profiles would call for similar screening responses within an actuarial framework. However, given the difference in risk pathways associated with Girl A and Girl B (e.g. higher rumination *v.* higher impairment) relying on a qualitative understanding of these cases may best inform subsequent clinical services based on the available resources.

plan to facilitate clinical decision-making, our findings provide a foundation for an evidence-based screening approach for prospectively forecasting first depression onset in youth.

Within translational research, it is not enough for an algorithm to be significant, but it also must be clinically useful. A traditional method for operationalizing clinical utility is by evaluating the positive and NPVs (see the bottom of Table 4) (Trevethan, 2017). In the present study, the PPVs were modest, suggesting that our algorithm is vulnerable to false-positive results (i.e. identifying youth at-risk for depression who do not go on to experience a first lifetime episode). In response, some may suggest abandoning a risk stratification approach for depression onset, and instead, use less selective preventive intervention strategies for depression or for clinicians to delineate idiographic risk factors via consultation with the patient (Carter *et al*., 2017; Large *et al*., 2017). Alternatively, because predictive values are correlated with base rates, others have argued that relying on these frequentist metrics may underestimate the utility of algorithms for outcomes with low base rates (e.g. 5–10%; Lavigne *et al*., 2016). Instead, EBM advocates recommend using posterior probabilities stemming from DLRs when evaluating clinical prediction models (Youngstrom, 2014). From this Bayesian perspective our algorithm was useful because it identified youth who were at 2–3 times elevated risk compared to the sample average, as well as individuals who were at minimal risk for experiencing depression onset. Thus, considering both perspectives, a reasonable conclusion when comparing our algorithm to current practices of using depressive symptoms is that we identified an incrementally valid approach for recommended preventive screening for depression onset (Siu, 2016). However, because of the low PPVs, our solution should not be considered a 'gold standard' and identifying an optimal strategy for predicting depression onset remains an important aim for future research. Thus, similar to existing algorithms for low base rate psychological outcomes (e.g. suicide; Fazel *et al*., 2019), it may be best to conceptualize our algorithm as a marker of continuous risk for depression onset as opposed to a strict classification tool.

Overall, our findings suggest that using rumination, impairment, *or* negative affect is a more reliable screening approach for depression onset compared to symptom inventories. Further, we found using the combination of indices provided incremental validity above and beyond using any measure independently. Using a multi-indicator approach facilitates a decision-making process that more closely aligns with the dimensional nature of adolescent depression (Hankin and Cohen, in press) and reduces the errors associated with making referrals based on scores near or at the cutoff on a single measure (Sheldrick *et al*., 2015). In addition, examining converging scores across rumination, impairment, and negative affect can inform evidence-based decision-making by understanding risk across multiple indicators. For instance, low impairment scores conferred DLRs below 0.25, suggesting that the onset of depression is highly unlikely with low scores on this particular index (Straus *et al*., 2011). Therefore, even with scores that approach, or even exceed, the cutoff for negative affect and rumination, one may not refer for services if these developmental risks are not manifesting in the context of impairment.

Ultimately, this study's algorithm can be tailored to a prevention protocol's resources so that evidence-based decisions can be made across a variety of settings. Table 5 helps to illustrate the clinical utility of our Cumulative Risk score. All three case examples represent actual screening profiles from 14-year-old girls who participated in the GEM study. As can be seen, the probability of prospectively being diagnosed with a first depression episode is approximately 50% higher than the sample average for those with risk scores of 5. Based on this probability, decisions can be made as to whether resources are available to engage these youth or if only those with a score of 6 (an over twofold risk for depression onset compared to the sample average) should be engaged. Alternatively, since each measure within the algorithm demonstrated a reliable and valid forecast for depression onset, *qualitative* decisions can also be made so that the screening process can best match the available prevention response. For

https://doi.org/10.1017/S0033291719002691 Published online by Cambridge University Press

instance, for 'Girl A' in Table 5, one of her high scores is rumination, while for 'Girl B' her rumination score is in the average range. If a cognitive–behavioral prevention program is offered in the area (e.g. Penn Resiliency Program; Gillham *et al.*, 2007), screening initiatives may prioritize 'Girl A' for a referral as it selects youth high in the mechanism of risk targeted by the preventative intervention (i.e. cognitive vulnerability). Thus, the algorithm retains flexibility through its dimensional risk approach, while still ensuring referrals and evidence-based decisions can be made to help connect youth to available services.

Strengths of the present study include the use of independent studies for development and validation, as well as repeated diagnostic interviews with relatively short recall periods to assess for prospective depression onset. At the same time, there are noteworthy limitations. First, it is important to replicate our findings outside of a research setting (Youngstrom *et al.*, 2017). These studies will be necessary for demonstrating the robust nature of our predictors across contexts and for identifying optimal cutoffs based on the objectives and resources of the setting, a necessary step for validating a screening algorithm (Youngstrom, 2014). Second, the current study used a pre-selection approach based on cross-validated univariate analyses to demonstrate which predictors may be useful for clinical decision-making (Steyerberg, 2009). However, other studies should replicate our findings using other analytic approaches (e.g. machine learning models), ideally with larger sample sizes across different length follow-ups, to determine if our findings generalize under different methodological conditions. Third, our study relied on subjective self- and parent-reported predictors for depression onset. Incorporating psychophysiological measures may demonstrate incremental improvements in predicting depression outcomes, particularly first lifetime episodes of depression (Cohen *et al.*, 2019). Fourth, the effect sizes and DLRs for our predictors were relatively small compared to screening tools for current outcomes (Youngstrom *et al.*, 2017). We suspect this reflects the difficulty in predicting prospective diagnostic outcomes compared to current diagnostic status. Although our effect sizes are comparable to other recent findings concerning prospective, depression-related outcomes (e.g. suicide; Fazel *et al.*, 2019) and first depression onset (Cohen *et al.*, 2019), future research should build upon our proposed algorithm to develop more targeted screening solutions for depression onset. Fifth, given the discontinuous nature of depression between childhood and adolescence (Cohen *et al.*, 2018*b*), it is important to examine if our algorithm can be extended downward to childhood. Alternatively, it should be tested if our findings generalize to detecting depression onset later in adolescence (e.g. ages 16–17), as the rates of depression markedly increase during this time (Twenge *et al.*, 2019). Ultimately, balancing personalized algorithms based on development and the translational benefits of having a 'one size fits all' approach has important clinical implications for depression prevention programming.

Finally, our recommended battery consisted of 58 items. This number is significantly lower than common mental health screening protocols (e.g. ASEBA assessments; Achenbach and Rescorla, 2001), yet it also comes at a time when briefer screening efforts are being encouraged (Lavigne *et al.*, 2016). Therefore, future studies could use Item Response Theory (IRT) to identify shorter versions of our screening battery (Youngstrom *et al.*, 2017). Continued efforts to include risk factor measures in preventative screening will fulfill the aims of universal pediatric depression screening (Siu, 2016), and ultimately, help reduce the prevalence and burden associated with depression onset at a young age (Kessler *et al.*, 2001; Weinberger *et al.*, 2017).

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291719002691.

Acknowledgements. This research was supported by NIMH grants (5R01MH077195, 5R01MH077178) awarded to Benjamin Hankin and Jami Young. Joseph Cohen's time on this manuscript was supported by the National Institute of Justice (2018-R2-CX-0022). The study protocol was approved by the institutional ethical review boards of the participating universities. All procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation with the Helsinki Declaration of 1975, as revised in 2008. Participants gave informed consent before participating. The authors have no conflicts of interest to report. JRC had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Financial support. This research was supported by the National Institute of Mental Health (NIMH) grants (5R01MH077195, 5R01MH077178) awarded to Benjamin Hankin and Jami Young. Joseph Cohen's time on this manuscript was supported by the National Institute of Justice (2018-R2-CX-0022).

References

Abela JRZ and Sullivan C (2003) A test of Beck's cognitive diathesis-stress theory of depression in early adolescents. *The Journal of Early Adolescence* **23**, 384–404.

Abela JRZ and Scheffler P (2008) Conceptualizing cognitive vulnerability to depression in youth: a comparison of the weakest link and additive approaches. *International Journal of Cognitive Therapy* **1**, 333–351.

Abela JRZ and Hankin BL (2011) Rumination as a vulnerability factor to depression during the transition from early to middle adolescence: a multiwave longitudinal study. *Journal of Abnormal Psychology* **120**, 259–271.

Abela JRZ, Vanderbilt E and Rochon A (2004) A test of the integration of the response styles and social support theories of depression in third and seventh grade children. *Journal of Social and Clinical Psychology* **23**, 653–674.

Achenbach TM and Rescorla LA (2001) *Manual for ASEBA School Age Forms & Profiles*. Burlington, VT: University of Vermont Research Center for Children, Youth, & Families.

Beck AT, Steer RA and Brown GK (1996) *Manual for the Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation.

Carter G, Milner A, McGill K, Pirkis J, Kapur N and Spittal MJ (2017) Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. *The British Journal of Psychiatry* **210**, 387–395.

Cohen JR, So FK, Hankin BL and Young JF (2018*a*) Translating cognitive vulnerability theory into improved adolescent depression screening: a receiver operating characteristic approach. *Journal of Clinical Child & Adolescent Psychology* **48**, 1–14. doi: 10.1080/15374416.2017.1416617.

Cohen JR, Andrews AR, Davis MM and Rudolph KD (2018*b*) Anxiety and depression during childhood and adolescence: testing theoretical models of continuity and discontinuity. *Journal of Abnormal Child Psychology* **46**, 1295–1308. https://doi.org/10.1007/s10802-017-0370-x.

Cohen JR, Thakur H, Burkhouse KL and Gibb BE (2019) A multimethod screening approach for pediatric depression onset: an incremental validity study. *Journal of Consulting and Clinical Psychology* **87**, 184–197.

Collins GS, Reitsma JB, Altman DG and Moons KG (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *Annals of Internal Medicine* **162**, 55–63.

Evans GW (2003) A multimethodological analysis of cumulative risk and allostatic load among rural children. *Developmental psychology* **39**, 924–933. http://dx.doi.org/10.1037/0012-1649.39.5.924.

Fabiano GA and Pelham Jr WE (2016) Impairment in children. In Goldstein S and Naglieri JA (eds), *Assessing Impairment*. New York, NY: Springer Press, pp. 71–79.

Farchione TJ, Fairholme CP, Ellard KK, Boisseau CL, Thompson-Hollands J, Carl JR, Gallagher MW and Barlow DH (2012) Unified protocol for transdiagnostic treatment of emotional disorders: a randomized controlled trial. *Behavior Therapy* **43**, 666–678.

Fazel S, Wolf A, Larrson H, Mallett S and Fanshawe TR (2019) The prediction of suicide in severe mental illness: development and validation of a clinical prediction rule (OxMIS). *Translational Psychiatry* **9**, 98.

Finkelhor D (2018) Screening for adverse childhood experiences (ACEs): cautions and suggestions. *Child Abuse & Neglect* **85**, 174–179.

Garber J (1984) The developmental progression of depression in female children. *New Directions for Child and Adolescent Development* **1984**, 29–58.

Garber J, Korelitz K and Samanez-Larkin S (2012) Translating basic psychopathology research to preventive interventions: a tribute to John R. Z. Abela. *Journal of Clinical Child & Adolescent Psychology* **41**, 666–681.

Gillham JE, Reivich KJ, Freres DR, Chaplin TM, Shatté AJ, Samuels B, Elkon AGL, Litzinger S, Lascher M, Gallop R and Seligman MEP (2007) School-based prevention of depressive symptoms: a randomized controlled study of the effectiveness and specificity of the Penn Resiliency Program. *Journal of Consulting and Clinical Psychology* **75**, 9–19.

Hankin BL and Abramson LY (2002) Measuring cognitive vulnerability to depression in adolescence: reliability, validity, and gender differences. *Journal of Clinical Child and Adolescent Psychology* **31**, 491–504.

Hankin BL and Cohen JR (in press) Child and adolescent depression. In Prinstein M and Youngstrom EA (eds), *Assessment of Childhood Disorders*. New York, NY: Guildford.

Hankin BL, Abramson LY, Moffitt TE, Silva PA, Mcgee R and Angell KE (1998) Development of depression from preadolescence to young adulthood: emerging gender differences in a 10-year longitudinal study. *Journal of Abnormal Psychology* **107**, 128–140.

Hankin BL, Stone L and Wright PA (2010) Corumination, interpersonal stress generation, and internalizing symptoms: accumulating effects and transactional influences in a multiwave study of adolescents. *Development and Psychopathology* **22**, 217–235.

Hankin BL, Young JF, Abela JRZ, Smolen A, Jenness JL, Gulley LD, Technow JR, Gottlieb AB, Cohen JR and Oppenheimer CW (2015) Depression from childhood into late adolescence: influence of gender, development, genetic susceptibility, and peer stress. *Journal of Abnormal Psychology* **124**, 803–816.

Hankin BL, Snyder HR and Gulley LD (2016) Cognitive risks in developmental psychopathology. In Cicchetti D (ed.), *Developmental Psychopathology*, 3rd Edn. Hoboken, NJ: Wiley Press, pp. 312–385.

Hanson R (2016) Assessing the calibration of actuarial risk scales. *Criminal Justice and Behavior* **44**, 26–39.

Jaycox L, Stein B, Paddock S, Miles JNV, Chandra A, Meredith LS, Tanielian T, Hickey S and Burnam MA (2009) Impact of teen depression on academic, social, and physical functioning. *Pediatrics* **124**, e596–e605.

Johnston C and Murray C (2003) Incremental validity in the psychological assessment of children and adolescents. *Psychological Assessment* **15**, 496–507.

Kaufman J, Birmaher B, Brent D, Rao U, Flynn C, Moreci P, Williamson D and Ryan N (1997) Schedule for affective disorders and schizophrenia for school-age children-present and lifetime version (K-SADS-PL): initial reliability and validity data. *Journal of the American Academy of Child & Adolescent Psychiatry* **36**, 980–988.

Kessler RC, Avenevoli S and Merikangas KR (2001) Mood disorders in children and adolescents: an epidemiologic perspective. *Biological Psychiatry* **49**, 1002–1014.

Klein DN, Dougherty LR and Olino TM (2005) Toward guidelines for evidence-based assessment of depression in children and adolescents. *Journal of Clinical Child and Adolescent Psychology* **34**, 412–432.

Kovacs M (1992) *Children's Depression Inventory* (*CDI*). Toronto, ON: Multi-Health Systems Inc.

Large MM, Ryan CJ, Carter G and Kapur N (2017) Can we usefully stratify patients according to suicide risk? *BMJ* **359**, 1–5. doi: doi.org/10.1136/bmj.j4627.

Laurent J, Catanzaro SJ, Joiner Jr TE, Rudolph KD, Potter KI, Lambert S, Osborne L and Gathright T (1999) A measure of positive and negative affect for children: scale development and preliminary validation. *Psychological Assessment* **11**, 326–338.

Lavigne JV, Meyers KM and Feldman M (2016) Systematic review: classification accuracy of behavioral screening measures for use in integrated primary care settings. *Journal of Pediatric Psychology* **41**, 1091–1109.

Lindhiem O, Petersen IT, Mentch LK and Youngstrom EA (2018) The importance of calibration in clinical psychology. *Assessment*, 1–15. doi: 10.1177/1073191117752055.

Lobbestael J, Leurgans M and Arntz A (2011) Inter-rater reliability of the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID I) and Axis II Disorders (SCID II). *Clinical Psychology & Psychotherapy* **18**, 75–79.

Monroe S and Harkness K (2011) Recurrence in major depression: a conceptual analysis. *Psychological Review* **118**, 655–674.

Muris P and Ollendick TH (2005) The role of temperament in the etiology of child psychopathology. *Clinical Child and Family Psychology Review* **8**, 271–289.

Nehmy T and Wade T (2015) Reducing the onset of negative affect in adolescents: evaluation of a perfectionism program in a universal prevention setting. *Behaviour Research and Therapy* **67**, 55–63.

Pettersson A, Boström KB, Gustavsson P and Ekselius L (2015) Which instruments to support diagnosis of depression have sufficient accuracy? A systematic review. *Nordic Journal of Psychiatry* **69**, 497–508.

Proctor MH, Moore LL, Gao D, Cupples LA, Bradlee ML, Hood MY and Ellison RC (2003) Television viewing and change in body fat from preschool to early adolescence: the Framingham Children's Study. *International Journal of Obesity* **27**, 827–833.

Sheldrick RC, Benneyan JC, Kiss IG, Briggs-Gowan MJ, Copeland W and Carter AS (2015) Thresholds and accuracy in screening tools for early detection of psychopathology. *Journal of Child Psychology and Psychiatry* **56**, 936–948.

Singer JD, Willett JB and Willett JB (2003) Applied longitudinal data analysis: Modeling change and event occurrence. Oxford university press.

Siu A (2016) Screening for depression in children and adolescents: U.S. Preventive Services Task Force recommendation statement. *Annals of Internal Medicine* **164**, 360–366.

Steyerberg E (2009) *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York: Springer.

Stockings E, Degenhardt L, Lee YY, Mihalopoulos C, Liu A, Hobbs M and Patton G (2015) Symptom screening scales for detecting major depressive disorder in children and adolescents: a systematic review and meta-analysis of reliability, validity and diagnostic utility. *Journal of affective disorders* **174**, 447–463. doi.org/10.1016/j.jad.2014.11.061

Straus SE, Glasziou P, Richardson WS and Haynes RB (2011) *Evidence-Based Medicine: How to Practice and Teach EBM*, 4th Edn. New York, NY: Churchill Livingstone.

Trevethan R (2017) Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. *Frontiers in Public Health* **5**, doi: 10.3389/fpubh.2017.00307.

Twenge JM, Cooper AB, Joiner TE, Duffy ME and Binau SG (2019) Age, period, and cohort trends in a nationally representative dataset, 2005-2017. *Journal of Abnormal Psychology* **128**, 185–199.

Weersing V, Jeffreys M, Do M, Schwartz K and Bolano C (2017) Evidence base update of psychosocial treatments for child and adolescent depression. *Journal of Clinical Child & Adolescent Psychology* **46**, 11–43.

Weinberger AH, Gbedemah M, Martinez AM, Nash D, Galea S and Goodwin RD (2017) Trends in depression prevalence in the USA from 2005 to 2015: widening disparities in vulnerable groups. *Psychological Medicine* **48**, 1308–1315.

Youngstrom EA (2014) A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: we are ready to ROC. *Journal of Pediatric Psychology* **39**, 204–221.

Youngstrom EA, Meter AV, Frazier TW, Hunsley J, Prinstein MJ, Ong ML and Youngstrom JK (2017) Evidence-based assessment as an integrative model for applying psychological science to guide the voyage of treatment. *Clinical Psychology: Science and Practice* **24**, 331–363.

Zuckerbrot R, Cheung A, Jensen P, Stein R and Laraque D (2018) Guidelines for adolescent depression in primary care (GLAD-PC): part I. Practice preparation, identification, assessment, and initial management. *Pediatrics* **141**, e20174081.