Introduction

The perinatal period, defined as the time from conception until one year after birth, is a high-risk time for the development of mental health disorders. Common mental disorders (CMDs) including depressive/mood disorders, post-traumatic stress disorder (PTSD), obsessive compulsive disorder (OCD) and anxiety disorders during the perinatal period account for a significant proportion of the global burden of disease in both high-income countries (HIC) and low- and middle-income countries (LMIC) (Gelaye et al., 2016; Howard et al., 2014). Prevalence estimates suggest up to 17% of the postnatal population have experienced depression (Shorey et al., 2018), with 2 in every 1,000 women requiring hospital admission for severe mental illness within the early postpartum period(Jones et al., 2014). Perinatal anxiety, bipolar affective disorder (BPAD) and PTSD also significantly contribute to maternal mortality and morbidity (Dennis et al., 2017; Micali et al., 2007; Wesseloo et al., 2016). The prevalence of clinically-diagnosed anxiety disorder has been estimated at 15.2% antenatally and 9.9% postnatally (Dennis 2017), with prevalence estimates ranging between 3.3-4.0% for perinatal PTSD (Yildiz 2017) and 0.7-11.1% for perinatal OCD (Leach 2017).

Various factors contribute to the risk of developing of CMDs in the perinatal period. Common risk factors include a previous history of psychopathology and psychosocial adversity, including low social support and abuse (Howard 2014). Additionally, childhood abuse and substance misuse alongside birth trauma and negative experience during pregnancy have been found to significantly increase the risk of developing CMDs in the perinatal period. This period of psychological vulnerability may not only contribute to relapses of pre-existing mental health conditions, but can precipitate the start of new onset psychiatric conditions such as postnatal depression, postnatal PTSD or, more rarely, puerperal psychosis.

Mental disorders during this critical period are not only associated with increased maternal mortality, suicide and self-harm(Howard and Khalifeh, 2020); data shows risks of adverse neonatal outcomes, such as foetal growth impairments, post-partum haemorrhage, placental abruption and stillbirths (McAllister-Williams et al., 2017; Rusner et al., 2016; Vigod et al., 2014), are greatly increased. Infants exposed to antenatal distress and children with ongoing exposure may also have both physical and psycho-social developmental issues including stunting, diarrhoeal infections and poor cognitive development(Gelaye et al., 2016). With an estimated one quarter of children exposed to maternal mental health disorders (Abel et al., 2019), timely identification and treatment of CMDs during the perinatal period is paramount.

In line with World Health Organisation (WHO) recommendations, if a condition is serious, prevalent and treatable, screening programmes should be implemented to identify high risk individuals ((WHO), 2020). Many screening tools have been developed to aid with the early identification and stratification of CMDs, with some specifically designed for perinatal women(Sit and Wisner, 2009) (e.g. Edinburgh Postnatal Depression Scale (EPDS)). Prior to use in clinical practice, validation of tools in the local context is essential to ensure appropriateness within the local population and establish context-specific cut-off thresholds. However, uncertainty regarding timing of implementation and appropriate cut-offs combined with low acceptability and inconsistent usage means many barriers to the detection of perinatal mental disorders remain (Milgrom and Gemmill, 2014; Rychnovsky and Brady, 2008; Sambrook Smith et al., 2019; Sit and Wisner, 2009). As a result, routine screening for perinatal CMDs, although supported by some countries, is not yet universally implemented (Milgrom and Gemmill, 2014).

Multiple systematic reviews summarising the validity of screening tools within the perinatal period exist. However, the breadth of data available, often spanning different population groups, countries and screening tools, means there remains a lack of certainty over which tools are suitable for use in perinatal populations. Therefore, a systematic review of these systematic reviews (referred to as an umbrella review throughout the rest of this paper) is

required to synthesise the evidence and provide clarifications where individual systematic reviews have not. The aim of this umbrella review was to provide an up-to-date summary of psychometric data of screening tools for the identification of perinatal CMDs. Our secondary aim was to explore the feasibility of validating these tools for use in clinical settings.

Methods

This review was pre-registered on Prospero [CRD42020199477] and reported according to *Preferred Reporting Items for Systematic Reviews and Meta-Analysis* (PRISMA) and Cochrane "Overviews of Reviews" guidance.

Search strategy

The search strategy was devised to retrieve all systematic reviews of validation studies of screening tools for perinatal CMDs. Relevant key search terms were identified through reviewing published literature within the field.

Search terms included the following keywords and MeSH terms: ("Perinatal" OR "Pregnancy" OR "Maternal Health") AND ("Mental Health" OR "Mental Disorder") AND ("Screening" OR "Validation" OR "Surveys and Questionnaires" OR "Checklist") AND ("Systematic Review" OR "Meta-Analysis") (full search strategy in supplementary file). The search strategy was adapted for use within the following electronic databases: Ovid MEDLINE, PsychINFO, EMBASE, Global Health and Cochrane Database of Systematic Reviews. It was run on June 1st 2020 and rerun on May 14th 2021. To identify further relevant papers and non-academic literature *Google Scholar* was searched and backwards citation searching of reference lists of included papers was conducted. Field experts were contacted to identify reviews underway or in press. Finally, we searched grey literature using BASE (Bielefeld Academic Search Engine), OpenGrey and WHO websites (including WHO International Maternal Mental Health Department).

We included systematic reviews which summarised evidence regarding the validation of screening tools used to identify CMDs during the perinatal period. The population was defined as any women of any age, in any trimester of pregnancy or up until twelve months post-partum. A systematic review was defined as having three discernible features: systematic collection of secondary data, critical appraisal of included studies and synthesis of collated findings. No language or publication date restrictions were set. Reviews assessing validation of tools in men were excluded. Studies were also excluded if they did not compare the screening tool against a clinical reference standard. Ideally this reference should be a diagnostic interview such as the Structured Clinical Interview for the Diagnosis of DSM Disorders (SCID-IV). However, we recognise that in some resource-poor settings, diagnostic interviews may not be feasible. Therefore, the tool had to be compared against what is currently considered 'best practice' in that setting, which may be another screening tool.

All papers returned by the search were imported into Covidence and duplicates removed. Titles and abstracts were screened and relevant papers were retrieved and reviewed in full. All full-texts were independently screened by two reviewers. A third reviewer screened a random selection of 20% of included studies to ensure suitability for inclusion. Where disagreement occurred regarding papers for inclusion a fourth reviewer was consulted.

Quality appraisal

To adopt an inclusive approach and gain breadth of knowledge, no minimal quality standard for included papers was set. Risk of bias was assessed using *A Measurement Tool to Assess Systematic Reviews* (AMSTAR-2)(Shea et al., 2017). AMSTAR-2 is a recognised

and validated tool for use in umbrella reviews and provides a clear and practical framework to critically appraise the quality of systematic reviews(Shea et al., 2017). A second reviewer independently appraised a random sample of 20% of papers. Included reviews were scored according to AMSTAR-2 criteria: “High” if no or one non-critical weakness; “Moderate” if more than one non-critical weakness; “Low” if one critical flaw; and “Critically Low” if more than one critical flaw. Agreement of scores was reached through consensus between two reviewers.

Data Analysis

Data extraction was completed by one reviewer with non-English papers translated and data extracted by a second author. Once input, data was re-reviewed and cleaned by a third author. Sensitivity and specificity for all tools at varying cut-offs was extracted. These data were available from individual systematic reviews either in the form of ranges or, where meta-analyses were conducted, pooled sensitivity and specificity with 95% confidence intervals (95% CI). When sensitivity and specificity was available for the same screening tool at the same cut-offs, data from individual reviews were compared using forest plots. All other results were summarised narratively using meta-synthesis.

Where possible, primary studies included within each systematic review were recorded. The degree of overlap between primary studies within each review was calculated using the Corrected Covered Area (CCA)(Hennessy and Johnson, 2020). The CCA between each pair of studies was presented as a pair plot with 0-5% considered slight overlap, 6-10% moderate overlap, 11-15% high overlap and >15% very high overlap. No studies were excluded based on CCA scores.

CCA= (N-r)/(rc-r)

c=Number of included reviews
r= number of publications of primary studies
N= number of total primary studies (including double counting)

Figure 1: formula for Corrected Covered Area (CCA) calculation

Role of the funding Source
There was no funding source for this review

Results

This search retrieved 7,891 studies. After removing duplicates, 6,841 studies were eligible for title and abstract screening. Of these, 340 full-texts were reviewed. Reasons for exclusion are listed in figure 2 (see appendix for reference list of excluded studies). In total, 31 papers commenting on 30 systematic reviews were included within this umbrella review (figure 2). Characteristics of included studies are shown in table 1 (appendix).

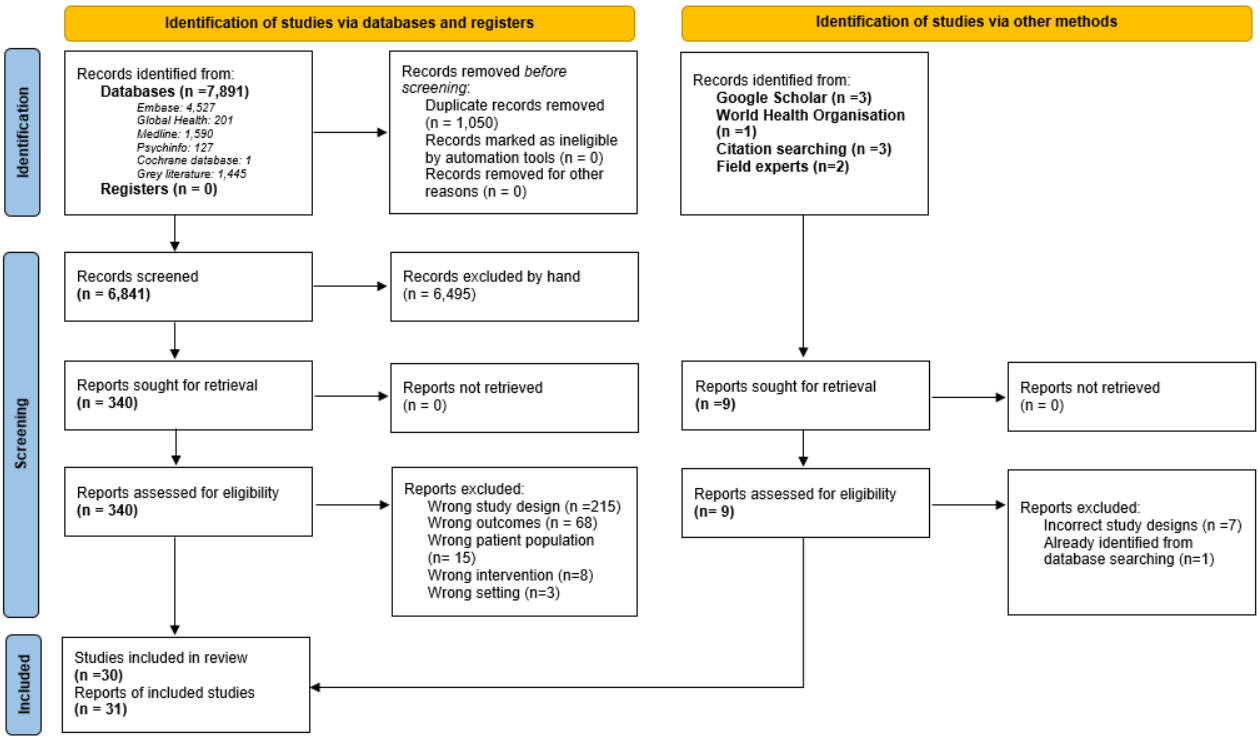


Figure 2: PRISMA flow chart of Study Selection

Description of Included Reviews

The number of individual studies in each review varied from 1 to 85 (mean number of studies per review = 7). Studies were published in English (n=29), French (n=1) and Portuguese (n=1). Included reviews were mostly conducted by American (n=11), English (n=8), Australian (n=2) and Canadian (n=2) institutions, but these reviews included studies from diverse geographical settings. All reviews were published between 2003 and 2021.

Reviews summarized data for between 1 and 21 tools each. In total, 76 different screening tools were identified with validity data only published for 43% (n=33) of these tools. Validity data was most commonly presented as sensitivity and specificity ranges at varying cut-offs. Seven reviews reported pooled sensitivity and specificity from meta-analysis (Chorwe-Sungani and Chipps, 2017; Fellmeth et al., 2021; Hewitt et al., 2009; Levis et al., 2020; Owora et al., 2016; Tsai et al., 2013a; Wang et al., 2020) and one review presented combined diagnostic odds ratios (Ali et al., 2016).

The most frequently-validated tools were the EPDS (n=28 reviews), Becks Depression Inventory (BDI) (n=13 reviews) and Patient Health Questionnaire (PHQ) (n=12 reviews) (figure 3). Of the 76 identified tools, 63 tools were validated in three or fewer reviews.

Twenty-seven reviews evaluated the validity of tools for depression, with most focused on postnatal depression (n=13) and two focused on antenatal depression. Four reviews validated tools for anxiety, most commonly perinatal anxiety (n=3). Other tools were validated for BPAD (n=1) and PTSD (n=1).

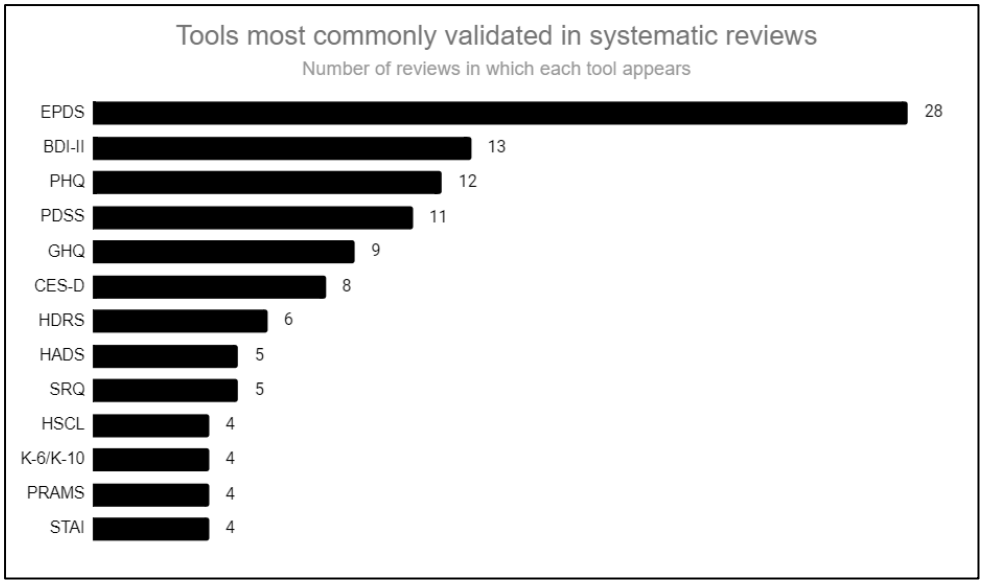


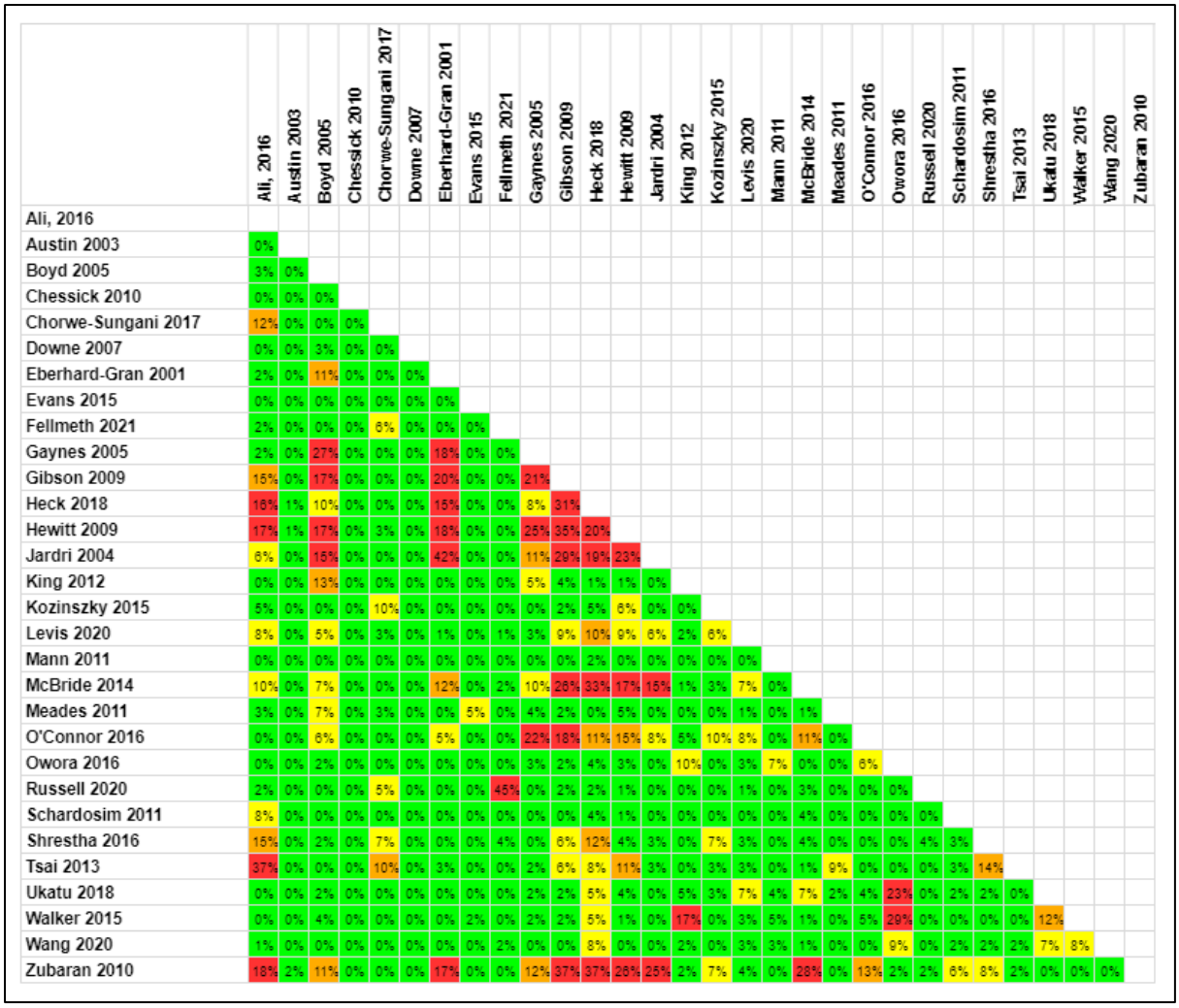
Figure 3: most commonly validated tools across all systematic reviews *Edinburgh Postnatal Depression Scale (EPDS); Beck Depression Inventory (BDI); Patient Health Questionnaire (PHQ); Postpartum Depression Screening Scale (PDSS); General Health Questionnaire (GHQ); The Centre for Epidemiological Studies - Depression scale (CES-D); Hamilton Depression Rating Scale (HDRS); Hospital Anxiety and Depression Scale (HADS); Self-Reporting Questionnaire (SRQ); Hopkins Symptom Checklist (HSCL); Kessler Psychological Distress Scale (K-6/K-10); Pregnancy Risk Assessment Monitoring System (PRAMS); State-Trait Anxiety Inventory (STAI)*

Quality of Included reviews

Using the AMSTAR-2 tool to asses quality, 54.8% (n=17) scored critically low and 25.8% (n=8) scored low. Reviews deemed to be high or moderate quality (n=6) conducted comprehensive literature searches, often examining three or more databases and provided detailed and replicable search strategies. Many systematic reviews used the QUADAS-2 tool to determine the quality of individual studies, but some reviews did not document use of any quality assessment or created their own risk of bias tool without justifications for this decision. This resulted in lower overall quality scores as authors did not discuss the impact of bias on meta-analyses or reasons for heterogeneity. Another risk of bias was that many reviews did not publish protocols *a priori*. Evidence from Cochrane Reviews suggests this is essential to prevent duplication in research and ensure stringent, transparent methodology.

Overlap Degree of Primary Studies

Out of the 31 systematic reviews included, there were a total 439 non-duplicate primary studies. The CCA was 2.98%; this “slight overlap” is likely due to variation in search strategies and multiple differing tools studied. The highest degree of overlap was between *Russell 2020*(Russell et al., 2020) and *Fellmeth 2021*(Fellmeth et al., 2021) with a CCA of 45% (figure 4). The most frequently included primary studies were *Beck 2001*(Beck and Gable, 2001) and *Garcia-Esteve 2003*(Garcia-Esteve et al., 2003) which were each included in 10 systematic reviews.



Ali, 2016	Austin 2003	Boyd 2005	Chessick 2010	Chorwe-Sungani 2017	Downe 2007	Eberhard-Gran 2001	Evans 2015	Fellmeth 2021	Gaynes 2005	Gibson 2009	Heck 2018	Hewitt 2009	Jardri 2004	King 2012	Kozinszky 2015	Levis 2020	Mann 2011	McBride 2014	Meades 2011	O'Connor 2016	Owora 2016	Russell 2020	Schardosim 2011	Shrestha 2016	Tsai 2013	Ukatu 2018	Walker 2015	Wang 2020	Zubaran 2010
Ali, 2016																													
Austin 2003	0%																												
Boyd 2005	3%	0%																											
Chessick 2010	0%	0%	0%																										
Chorwe-Sungani 2017	12%	0%	0%	0%																									
Downe 2007	0%	0%	3%	0%	0%																								
Eberhard-Gran 2001	2%	0%	11%	0%	0%	0%																							
Evans 2015	0%	0%	0%	0%	0%	0%	0%																						
Fellmeth 2021	2%	0%	0%	0%	6%	0%	0%	0%																					
Gaynes 2005	2%	0%	27%	0%	0%	0%	18%	0%	0%																				
Gibson 2009	15%	0%	17%	0%	0%	0%	20%	0%	0%	21%																			
Heck 2018	16%	1%	10%	0%	0%	0%	15%	0%	0%	8%	31%																		
Hewitt 2009	17%	1%	17%	0%	3%	0%	18%	0%	0%	25%	35%	20%																	
Jardri 2004	6%	0%	15%	0%	0%	0%	42%	0%	0%	11%	29%	19%	23%																
King 2012	0%	0%	13%	0%	0%	0%	0%	0%	0%	5%	4%	1%	1%	0%															
Kozinszky 2015	5%	0%	0%	0%	10%	0%	0%	0%	0%	0%	2%	5%	6%	0%	0%														
Levis 2020	8%	0%	5%	0%	3%	0%	1%	0%	1%	3%	9%	10%	9%	6%	2%	6%													
Mann 2011	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	0%	0%	0%	0%	0%												
McBride 2014	10%	0%	7%	0%	0%	0%	12%	0%	2%	10%	26%	33%	17%	15%	1%	3%	7%	0%											
Meades 2011	3%	0%	7%	0%	3%	0%	0%	5%	0%	4%	2%	0%	5%	0%	0%	0%	1%	0%	1%										
O'Connor 2016	0%	0%	6%	0%	0%	0%	5%	0%	0%	22%	18%	11%	15%	8%	5%	10%	8%	0%	11%	0%									
Owora 2016	0%	0%	2%	0%	0%	0%	0%	0%	0%	3%	2%	4%	3%	0%	10%	0%	3%	7%	0%	0%	6%								
Russell 2020	2%	0%	0%	0%	5%	0%	0%	0%	45%	0%	2%	2%	1%	0%	0%	0%	1%	0%	3%	0%	0%	0%							
Schardosim 2011	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	1%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%						
Shrestha 2016	15%	0%	2%	0%	7%	0%	0%	0%	4%	0%	6%	12%	4%	3%	0%	7%	3%	0%	4%	0%	0%	0%	4%	3%					
Tsai 2013	37%	0%	0%	0%	10%	0%	3%	0%	0%	2%	6%	8%	11%	3%	0%	3%	3%	0%	1%	9%	0%	0%	0%	3%	14%				
Ukatu 2018	0%	0%	2%	0%	0%	0%	0%	0%	0%	2%	2%	5%	4%	0%	5%	3%	7%	4%	7%	2%	4%	23%	0%	2%	2%	0%			
Walker 2015	0%	0%	4%	0%	0%	0%	0%	2%	0%	2%	2%	5%	1%	0%	17%	0%	3%	5%	1%	0%	8%	29%	0%	0%	0%	0%	12%		
Wang 2020	1%	0%	0%	0%	0%	0%	0%	0%	2%	0%	0%	8%	0%	0%	2%	0%	3%	3%	1%	0%	0%	9%	0%	2%	2%	2%	7%	8%	
Zubaran 2010	18%	2%	11%	0%	0%	0%	17%	0%	0%	12%	37%	37%	26%	25%	2%	7%	4%	0%	28%	0%	13%	2%	2%	6%	8%	2%	0%	0%	0%

Colour	Interpretation
	Slight overlap (≤5%)
	Moderate overlap (5.1-10%)
	High overlap (10.1-15%)
	Very high overlap (≥15.1%)

Figure 4: Overlap matrix providing Corrected Covered Area percentages between each pair of systematic reviews *(Note both reviews written by Owora have been combined into a single column as each publication makes comments based on the same sample of primary studies).*

Assessment tools

Validity scores for tools reported in 12 or more reviews were summarised using forest plots of pooled estimates with 95% confidence intervals or ranges. Sufficient data was available for the EPDS, BDI and PHQ (figure 5-6 and supplementary data).Forest plots were stratified according to timing of application (antenatal, postnatal or perinatal) and presented in order of tool cut-off scores. A pattern of decreasing sensitivity and increasing specificity with increase in cut-off of the EPDS score was observed. Although there were fewer estimates available and 95% CIs and ranges were wider, this pattern was also observed in the BDI and PHQ tools (see supplementary data).

Figure 5: Forest plot of EPDS sensitivity stratified by time period of tool application

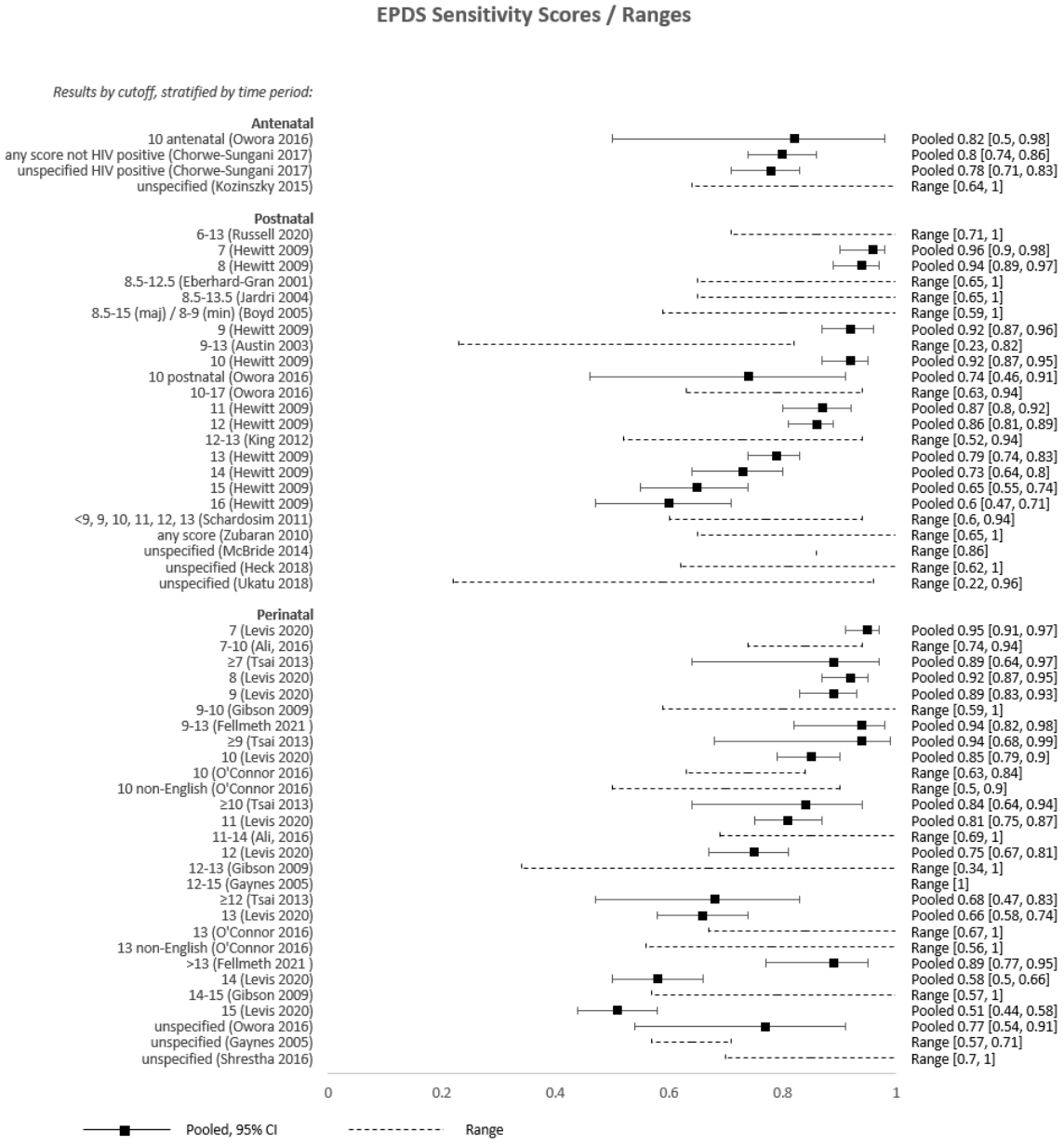
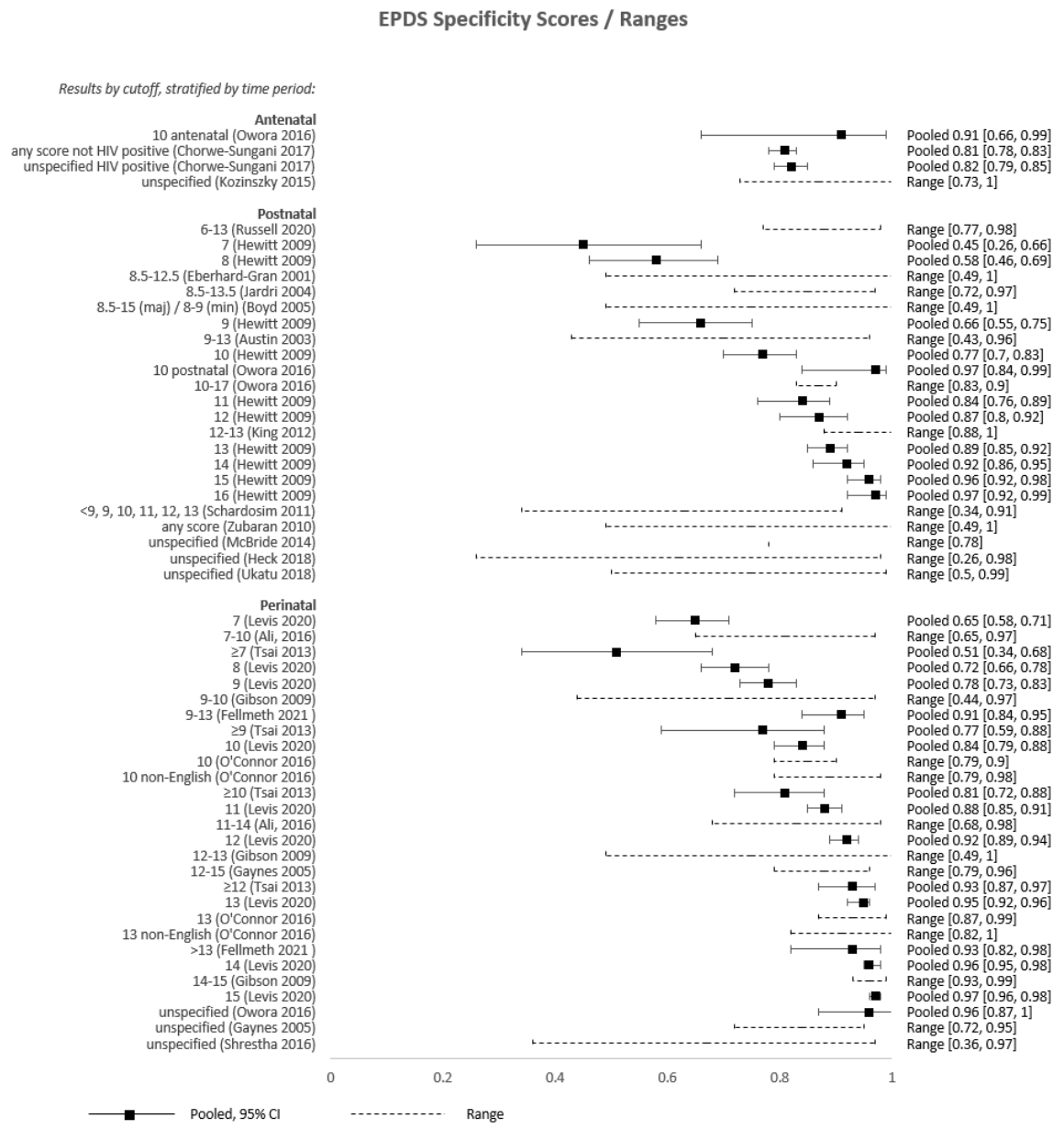
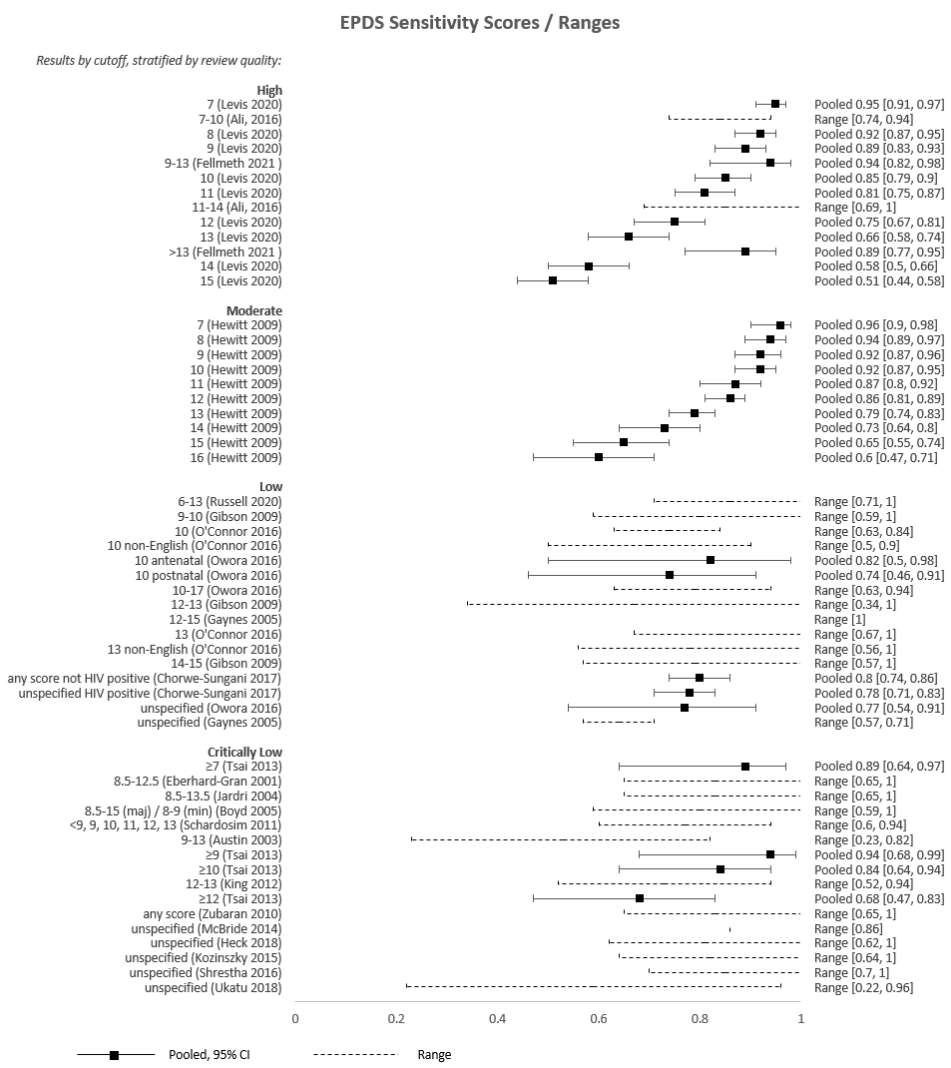


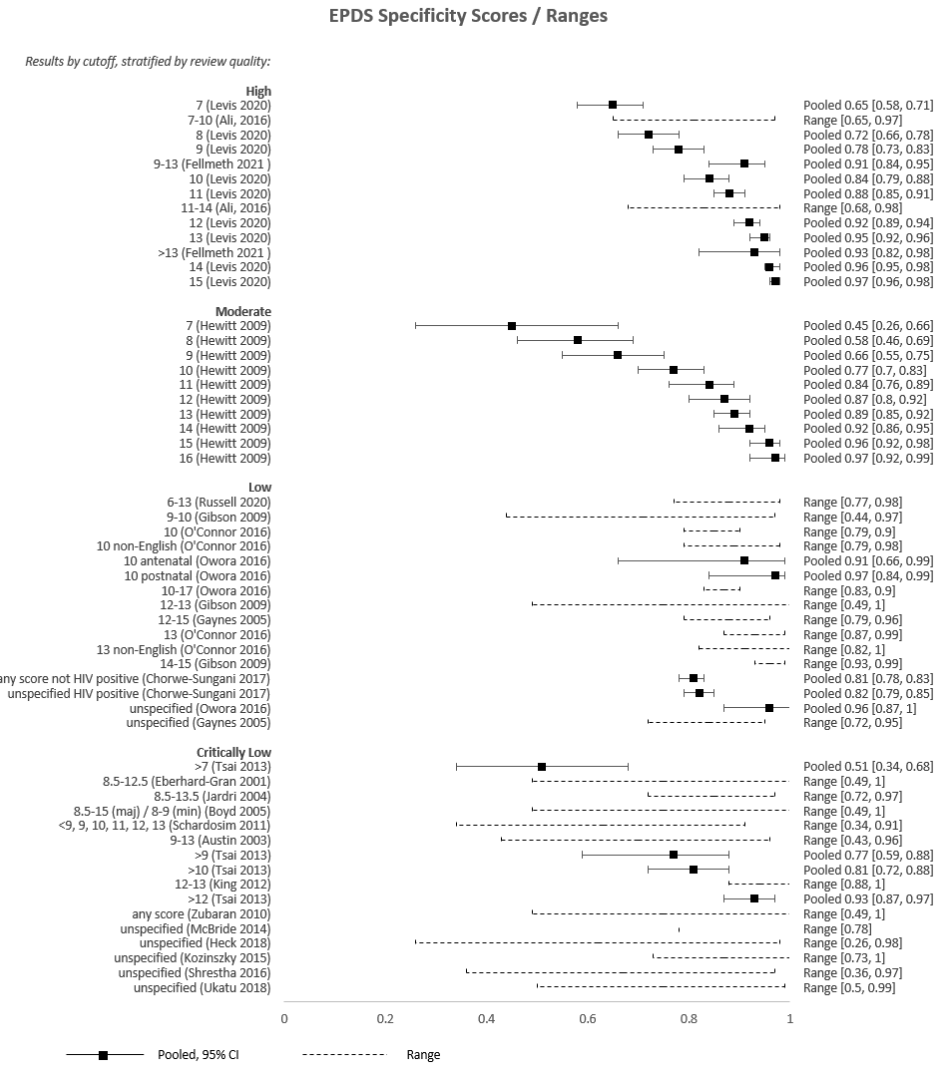
Figure 6: Forest plot of EPDS specificity stratified by time period of tool application



We also explored sensitivity and specificity of the EPDS according to study quality scores (figure 7). This showed similar patterns of increasing specificity with increasing cut-offs in high and moderate quality papers. This pattern was more variable amongst low and critically low-quality papers, where pooled estimates had wider 95% CIs and overall specificity was lower.

Figure 7: Forest plot of validity data for EPDS stratified by quality scores





Context-specific Application of Tools

The included reviews discussed studies from a diverse range of geographical, cultural and socio-economic settings. Two reviews highlighted that validity varied according to socio-economic background and therefore validation according to income and socio-economic background may be important (Chorwe-Sungani and Chipps, 2017; King, 2012). One review reported that sensitivity and specificity were higher among more affluent populations relative to those from more deprived backgrounds (King, 2012). Two papers provided evidence supporting the use of the BDI, EPDS and Postpartum Depression Screening Scale (PDSS) for identifying postpartum depression in women living in poverty (Chorwe-Sungani and Chipps, 2017; King, 2012).

Seven reviews highlighted the importance of ethnicity (Boyd et al., 2005; Gibson et al., 2009; Heck, 2018; King, 2012; Kozinszky and Dudas, 2015; Owora et al., 2016; Zubaran et al., 2010a). One review suggested different cut-offs should be used for different cultural groups(Zubaran et al., 2010b). This was supported by another review which found that optimal sensitivity and specificity across racial and ethnic groups could only be achieved by adjusting cut-off scores for different groups of women(King, 2012). This was also reflected in one review as higher cut-offs did not always produce higher sensitivity or specificity, likely due to cultural differences(Kozinszky and Dudas, 2015). One review suggested that the disparity may be because of the “test taking culture” in Western contexts compared to women in other cultures and settings who may find the experience less familiar, resulting in reduced validity scores(Downe et al., 2007).

The importance of literacy, cross-cultural translations of tools and feasibility of using screening tools developed in English and translated into other languages was also often discussed(Boyd et al., 2005; Fellmeth et al., 2021; McBride et al., 2014; Walker et al., 2015). Two reviews suggested that there should be higher cut-offs for Vietnamese- vs. Arabic-speaking populations in Australia(Boyd et al., 2005; Gibson et al., 2009). Despite this, one review looking specifically into the use of idiomatic language, forms of expression and the impact on psychometric statistical analysis did not find any examples of where this affected tool application(Walker et al., 2015). One paper concluded that if a range of languages is spoken within a population then local translations of the EPDS were the most suitable tool for depression screening in clinical practice(Boyd et al., 2005).

Three reviews hypothesized the possible impact of mothers’ age on validity and identification of CMDs(Chorwe-Sungani and Chipps, 2017; Fellmeth et al., 2021; McBride et al., 2014). One review highlighted that EPDS sensitivity was 5% lower in teenage mothers compared to adult pregnant women(Chorwe-Sungani and Chipps, 2017).

The ideal timing of tool application may vary between individuals and according to the context of service delivery. Sensitivity and specificity varied according to the time within the perinatal period that they were applied. Four reviews suggested that the immediate postnatal period and first trimester were, despite being high-risk periods, the most problematic for screening (Meades and Ayers, 2011; Owora et al., 2016; Schardosim and Heldt, 2011; Shrestha et al., 2016). Although screening often occurred within a few days postpartum, one review found that this was when the majority of tools had the lowest sensitivity for anxiety diagnosis(Meades and Ayers, 2011). Three reviews highlighted higher diagnostic performance of instruments during the second and third trimesters(Owora et al., 2016; Rusner et al., 2016; Tsai et al., 2013a). However in BPAD, screening had highest sensitivity if performed during the first antenatal visit and within the first few days postpartum(Chessick and Dimidjian, 2010).

Feasibility Considerations for Tool Application

Several feasibility considerations and the impact on validity were considered within the included reviews. One review found that tools with fewer than 25 questions were more practical to administer(Chessick and Dimidjian, 2010) and that those which took an average of five minutes to complete were more acceptable to women than more time-consuming tools(Fellmeth et al., 2021). One review highlighted a pattern of lower specificity but higher sensitivity with fewer items per screening tool(Owora et al., 2016). Two studies reported that delivery method could affect psychometric properties(Eberhard-Gran et al., 2001; Jardri, 2004). One study suggested that women preferred to talk about how they felt to an interviewer rather than filling out questionnaires(Jardri, 2004). However another study found that self-reported measures had significantly higher estimates of postpartum depression compared to interview-based delivery(Eberhard-Gran et al., 2001).

Discussion

Our research aimed to identify which screening tools for the identification of CMDs have been validated for use in the perinatal period. This umbrella review identified 31 systematic reviews (from 439 non-duplicate primary studies) covering 76 individual tools.

Despite this vast number of tools identified, we found no evidence to suggest that any one tool was “more valid” than another. For tools with sufficient data, the observed pattern of sensitivity and specificity relative to cut-off scores was as expected: decreasing sensitivity and increasing specificity with increasing cut-offs. The EPDS was the most frequently validated tool with good evidence to support use in depression across the perinatal period. There was also some evidence for the use of tools which are not specific to the perinatal period such as BDI and PHQ. Although these tools were not designed with perinatal women in mind, the evidence shows that both the BDI and PHQ can be valid for use in perinatal women.

Most of the tools were validated for use in depression. Gaps in the evidence exist for the validation of anxiety measures. Lacking data on anxiety prevented direct comparison of anxiety measures and validity data across systematic reviews, therefore uncertainty remains regarding the validity of anxiety screening tools. The current literature highlights challenges associated with defining and measuring perinatal anxiety(Harrison and Alderdice, 2020). Screening tools tend to identify a set of core symptoms relating to a variety of anxiety disorders (e.g. panic disorder or phobias). These disorders have varying aetiology, triggers and manifestations – nuances which can be elicited in diagnostic interviews but are often missed by screening tools. Therefore, anxiety screening results can be difficult to interpret. Some tools which are primarily designed to identify a single diagnosis (e.g. antenatal depression and the Whooley questions) may also identify other CMDs (e.g. perinatal anxiety)(Howard et al., 2018). For example, the EPDS composes of 10 closed questions designed to identify likely postnatal depression. However, question 4 and 5 focus on feeling anxious and scared/panicky so, although this tool is widely used for postnatal depression, this may too be identifying perinatal anxiety. Tools designed to identify symptomatology of distress rather than a specific perinatal pathology (e.g. *Kessler Psychological Distress Scale* and *General Health Questionnaire*) may be useful as alternative primary screening tools where it is not possible to screen for every individual CMD.

Our review found limited validity data on the use of tools in the most high-risk perinatal periods. We found limited and varying evidence for the antenatal period. For example, only

three systematic reviews provided evidence on the sensitivity and specificity of the EPDS antenatally with wide ranging results. Most reported data were from the late postnatal period, with qualitative data highlighting difficulties of tool application in the early antenatal and immediate postnatal period. Reduced validity of the tools during these times is expected; rapid hormonal variations coupled with socio-environmental changes make these high-risk periods for development of CMDs. Therefore, results of screening during early antenatal and immediate postnatal periods must be interpreted with caution. For example, there may be a higher false positive rate of depression when screening tools are applied in the immediate postpartum period due the concurrence of “baby blues” (Degner, 2017). Where possible, repeated measurements over time may help to ensure correct identification of women with CMD.

The context in which screening tools are applied is perhaps the most important factor determining validity. This review has found that some of the highest validity and best quality research comes from context-specific validation of tools. Context not only refers to geographical location but also socio-economic background, ages, language, educational status and cultural background of women in the target population. Recent literature also suggests that conditions under which the screening tool is applied should be carefully considered to ensure that high-risk women can be adequately followed-up and supported (Hirshler et al., 2021). Importantly, the local validation of screening tools avoids Anglo-centric and westernized ideals of CMDs from being applied to women in diverse settings.

Subgroup analysis of EPDS data according to quality scores showed that pooled sensitivity and specificity estimates were very similar irrespective of our objective quality scores. We found that quality scores did not differentiate results. Despite wider confidence intervals in the lower quality studies the pooled sensitivities were similar at most cut-offs (Austin and Lumley, 2003; Hewitt et al., 2009; Levis et al., 2020; Tsai et al., 2013b; Ukatu et al., 2019). Only four papers scored high quality with the AMSTAR 2 tool (Ali et al., 2016; Downe et al., 2007; Fellmeth et al., 2021; Levis et al., 2020). These reviews had pre-registered protocols, documented their excluded studies and used established risk of bias tools. The AMSTAR tool weights these factors as critical to the identification of risk of bias. We considered this weighting of criteria to be disproportionate, favoring transparency of reporting rather than methodological quality, and so quality scores were lower than we expected for reviews. The ROBINS tool, although not specific to systematic reviews, offers an alternative quality appraisal tool for consideration as it focuses on additional domains such as missing data bias and preferential reporting of results(Sterne et al., 2016).

A strength of this review is the broad nature of the search; this allowed for inclusion of diverse studies published in different languages and regions. A weakness of this umbrella review is that despite a low CCA score, overlap of primary studies still remained; therefore, we were still unable perform meta-analysis to provide pooled sensitivity and specificity estimates for each tool. Another possible limitation of this review is that we were unable to collect data on the mode of administration of screening tools (e.g. by healthcare workers or researchers, verbally-administered or self-response) in the 439 included primary studies, which may explain some of the heterogeneity between results. To maximize validity, screening tools should ideally be administered by those with appropriate training and skills. In this review we also aimed to explore feasibility considerations for validating screening tools in clinical settings; however only few of our included studies reported on this outcome. To allow for more in-depth discussion on perceived acceptability and barriers to tool application a more targeted search methodology with an emphasis on identifying qualitative studies could be used.

In summary, this comprehensive umbrella review identified 76 screening tools for identification of CMDs in the perinatal period. The EPDS, BDI and PHQ were the most commonly-validated tools and evidence suggests they are valid across a range of diverse settings and perinatal populations. Large evidence gaps exist in validation of screening tools for anxiety, PTSD and BPAD, in high-risk groups and at high-risk time periods (e.g. the early antenatal and immediate postnatal period). This umbrella review highlights that utilizing screening tools in clinical practice is complex and requires careful consideration of the population, psychological, social and cultural risk factors and the wider context and health system.

Supplementary Material

Figure 1: Medline Search Strategy

Table 1: Characteristics of included systematic reviews

Table 2: Reference list of tools: ID numbers and abbreviation key

Table 3: Reference list of tools: frequency of validation and review reference

Figure 2: Forest plots for sensitivity and specificity estimates of BDI

Figure 3: Forest plots of sensitivity and specificity estimates of PHQ

Table 4: Breakdown of AMSTAR-2 Quality scores

Appendix file: Excluded studies at full-text screening stage

References

(WHO), W.H.O., 2020. Screening Programmes: a short guide to increase effectiveness, maximize benefits and minimize harm. WHO Regional Office for Europe, Copenhagen.
Abel, K.M., Hope, H., Swift, E., Parisi, R., Ashcroft, D.M., Kosidou, K., Osam, C.S., Dalman, C., Pierce, M., 2019. Prevalence of maternal mental illness among children and adolescents in the UK between 2005 and 2017: a national retrospective cohort analysis. Lancet Public Health 4, e291-e300.
Ali, G.C., Ryan, G., De Silva, M.J., 2016. Validated Screening Tools for Common Mental Disorders in Low and Middle Income Countries: A Systematic Review. PLoS One 11, e0156939.
Austin, M., Lumley, J., 2003. Antenatal screening for postnatal depression: A systematic review. Acta Psychiatrica Scandinavica 107, 10-17.
Beck, C.T., Gable, R.K., 2001. Comparative analysis of the performance of the Postpartum Depression Screening Scale with two other depression instruments. Nurs Res 50, 242-250.
Boyd, R.C., Le, H., Somberg, R., 2005. Review of screening instruments for postpartum depression. Archives of Women's Mental Health 8, 141-153.
Chessick, C.A., Dimidjian, S., 2010. Screening for bipolar disorder during pregnancy and the postpartum period. Arch Womens Ment Health 13, 233-248.
Chorwe-Sungani, G., Chipps, J., 2017. A systematic review of screening instruments for depression for use in antenatal services in low resource settings. BMC Psychiatry 17, 112.

Degner D. Differentiating between “baby blues,” severe depression, and psychosis. Bmj. 2017 Nov 10;359.
Dennis, C.L., Falah-Hassani, K., Shiri, R., 2017. Prevalence of antenatal and postnatal anxiety: systematic review and meta-analysis. Br J Psychiatry 210, 315-323.
Downe, S.M., Butler, E., Hinder, S., 2007. Screening tools for depressed mood after childbirth in UK-based South Asian women: a systematic review. J Adv Nurs 57, 565-583.
Eberhard-Gran, M., Eskild, A., Tambs, K., Opjordsmoen, S., Samuelsen, S., 2001. Review of validation studies of the Edinburgh Postnatal Depression Scale. Acta psychiatrica Scandinavica 104, 243-249.
Fellmeth, G., Harrison, S., Opondo, C., Nair, M., Kurinczuk, J.J., Alderdice, F., 2021. Validated screening tools to identify common mental disorders in perinatal and postpartum women in India: a systematic review and meta-analysis. BMC psychiatry 21, 1-10.
Garcia-Esteve, L., Ascaso, C., Ojuel, J., Navarro, P., 2003. Validation of the Edinburgh Postnatal Depression Scale (EPDS) in Spanish mothers. Journal of Affective Disorders 75, 71-76.
Gelaye, B., Rondon, M.B., Araya, R., Williams, M.A., 2016. Epidemiology of maternal depression, risk factors, and child outcomes in low-income and middle-income countries. The lancet. Psychiatry 3, 973-982.
Gibson, J., McKenzie-Mcharg, K., Shakespeare, J., Price, J., Gray, R., 2009. A systematic review of studies validating the Edinburgh Postnatal Depression Scale in antepartum and postpartum women. Acta Psychiatrica Scandinavica 119, 350-364.
Harrison, S., Alderdice, F., 2020. Challenges of defining and measuring perinatal anxiety. Journal of Reproductive and Infant Psychology 38, 1-2.
Heck, J.L., 2018. Screening for postpartum depression in American Indian/Alaska Native women: A comparison of two instruments. American Indian and Alaska Native Mental Health Research 25, 74-102.
Hennessy, E.A., Johnson, B.T., 2020. Examining overlap of included studies in meta-reviews: Guidance for using the corrected covered area index. Research synthesis methods 11, 134-145.
Hewitt, C.E., Gilbody, S.M., Brealey, S., Paulden, M., Palmer, S., Mann, R., Green, J., Morrell, J., Barkham, M., Light, K., Richards, D., 2009. Methods to identify postnatal depression in primary care: An integrated evidence synthesis and value of information analysis. Health Technology Assessment 13, iii-206.
Hirshler Y, Gemmill AW, Milgrom J. An Australian perspective on treating perinatal depression and anxiety: a brief review of efficacy and evidence-based practice in screening, psychosocial assessment and management. Annali dell'Istituto Superiore di Sanità. 2021 Mar 25;57(1):40-50
Howard, L.M., Khalifeh, H., 2020. Perinatal mental health: a review of progress and challenges. World psychiatry : official journal of the World Psychiatric Association (WPA) 19, 313-327.
Howard, L.M., Molyneaux, E., Dennis, C.L., Rochat, T., Stein, A., Milgrom, J., 2014. Non-psychotic mental disorders in the perinatal period. Lancet 384, 1775-1788.
Howard, L.M., Ryan, E.G., Trevillion, K., Anderson, F., Bick, D., Bye, A., Byford, S., O'Connor, S., Sands, P., Demilew, J., Milgrom, J., Pickles, A., 2018. Accuracy of the Whooley questions and the Edinburgh Postnatal Depression Scale in identifying depression and other mental disorders in early pregnancy. Br J Psychiatry 212, 50-56.
Jardri, R., 2004. Postnatal depression screening: A Qualitative review of the validation studies of the Edinburgh Postnatal Depression Scale. Devenir 16, 246-262.
Jones, I., Chandra, P.S., Dazzan, P., Howard, L.M., 2014. Bipolar disorder, affective psychosis, and schizophrenia in pregnancy and the post-partum period. Lancet 384, 1789-1799.
King, P.A.L., 2012. Validity of postpartum depression screening across socioeconomic groups: A review of the construct and common Screening tools. Journal of Health Care for the Poor and Underserved 23, 1431-1456.
Kozinszky, Z., Dudas, R.B., 2015. Validation studies of the Edinburgh Postnatal Depression Scale for the antenatal period. Journal of Affective Disorders 176, 95-105.

Leach LS, Poyser C, Cooklin AR, Giallo R. Prevalence and course of anxiety disorders (and symptom levels) in men across the perinatal period: A systematic review. J Affect Disord. 2016 Jan 15;190:675-686. doi: 10.1016/j.jad.2015.09.063.
Levis, B., Negeri, Z., Sun, Y., Benedetti, A., Thombs, B.D., 2020. Accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for screening to detect major depression among pregnant and postpartum women: systematic review and meta-analysis of individual participant data. bmj 371.
McAllister-Williams, R.H., Baldwin, D.S., Cantwell, R., Easter, A., Gilvarry, E., Glover, V., Green, L., Gregoire, A., Howard, L.M., Jones, I., Khalifeh, H., Lingford-Hughes, A., McDonald, E., Micali, N., Pariante, C.M., Peters, L., Roberts, A., Smith, N.C., Taylor, D., Wieck, A., Yates, L.M., Young, A.H., 2017. British Association for Psychopharmacology consensus guidance on the use of psychotropic medication preconception, in pregnancy and postpartum 2017. J Psychopharmacol 31, 519-552.
McBride, H.L., Wiens, R.M., McDonald, M.J., Cox, D.W., Chan, E.K.H., 2014. The Edinburgh Postnatal Depression Scale (EPDS): A review of the reported validity evidence, Validity and validation in social, behavioral, and health sciences. Springer International Publishing; Switzerland, Cham, Switzerland, pp. 157-174.
Meades, R., Ayers, S., 2011. Anxiety measures validated in perinatal populations: a systematic review. J Affect Disord 133, 1-15.
Micali, N., Treasure, J., Simonoff, E., 2007. Eating disorders symptoms in pregnancy: a longitudinal study of women with recent and past eating disorders and obesity. J Psychosom Res 63, 297-303.
Milgrom, J., Gemmill, A.W., 2014. Screening for perinatal depression. Best Practice & Research in Clinical Obstetrics & Gynaecology 28, 13-23.
Owora, A.H., Carabin, H., Reese, J., Garwe, T., 2016. Summary diagnostic validity of commonly used maternal major depression disorder case finding instruments in the United States: A meta-analysis. Journal of Affective Disorders 205, 335-343.
Rusner, M., Berg, M., Begley, C., 2016. Bipolar disorder in pregnancy and childbirth: a systematic review of outcomes. BMC Pregnancy Childbirth 16, 331.
Russell, P.S.S., Chikkala, S.M., Earnest, R., Viswanathan, S.A., Russell, S., Mammen, P.M., 2020. Diagnostic accuracy and clinical utility of non-English versions of Edinburgh Post-Natal Depression Scale for screening post-natal depression in India: A meta-analysis. World Journal of Psychiatry 10, 71-80.
Rychnovsky, J.D., Brady, M.A., 2008. Choosing a postpartum depression screening instrument for your pediatric practice. Journal of Pediatric Health Care 22, 64-67.
Sambrook Smith, M., Lawrence, V., Sadler, E., Easter, A., 2019. Barriers to accessing mental health services for women with perinatal mental illness: systematic review and meta-synthesis of qualitative studies in the UK. BMJ Open 9, e024803.
Schardosim, J.M., Heldt, E., 2011. [Postpartum depression screening scales: a systematic review]. Revista Gaucha de Enfermagem 32, 159-166.
Shea, B.J., Reeves, B.C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., Henry, D.A., 2017. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 358, j4008.
Shorey, S., Chee, C.Y.I., Ng, E.D., Chan, Y.H., Tam, W.W.S., Chong, Y.S., 2018. Prevalence and incidence of postpartum depression among healthy mothers: A systematic review and meta-analysis. J Psychiatr Res 104, 235-248.
Shrestha, S.D., Pradhan, R., Tran, T.D., Gualano, R.C., Fisher, J.R.W., 2016. Reliability and validity of the Edinburgh Postnatal Depression Scale (EPDS) for detecting perinatal common mental disorders (PCMDs) among women in low-and lower-middle-income countries: A systematic review. BMC Pregnancy and Childbirth 16.
Sit, D.K., Wisner, K.L., 2009. Identification of postpartum depression. Clinical Obstetrics & Gynecology 52, 456-468.
Sterne, J.A.C., Hernán, M.A., Reeves, B.C., Savović, J., Berkman, N.D., Viswanathan, M., Henry, D., Altman, D.G., Ansari, M.T., Boutron, I., Carpenter, J.R., Chan, A.-W., Churchill, R., Deeks, J.J.,

Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y.K., Pigott, T.D., Ramsay, C.R., Regidor, D., Rothstein, H.R., Sandhu, L., Santaguida, P.L., Schünemann, H.J., Shea, B., Shrier, I., Tugwell, P., Turner, L., Valentine, J.C., Waddington, H., Waters, E., Wells, G.A., Whiting, P.F., Higgins, J.P.T., 2016. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 355, i4919.
Tsai, A.C., Scott, J.A., Hung, K.J., Zhu, J.Q., Matthews, L.T., Psaros, C., Tomlinson, M., 2013a. Reliability and validity of instruments for assessing perinatal depression in African settings: Systematic review and meta-analysis. PLoS ONE 8.
Tsai, A.C., Scott, J.A., Hung, K.J., Zhu, J.Q., Matthews, L.T., Psaros, C., Tomlinson, M., 2013b. Reliability and validity of instruments for assessing perinatal depression in African settings: systematic review and meta-analysis. PLoS One 8, e82521.
Ukatu, N., Brulja, M., Clare, C.A., 2019. Postpartum depression screening tools: A review. Journal of Women's Health 28 (6), 10.
Vigod, S.N., Kurdyak, P.A., Dennis, C.L., Gruneir, A., Newman, A., Seeman, M.V., Rochon, P.A., Anderson, G.M., Grigoriadis, S., Ray, J.G., 2014. Maternal and newborn outcomes among women with schizophrenia: a retrospective population-based cohort study. Bjog 121, 566-574.
Walker, L.O., Gao, J., Xie, B., 2015. Postpartum Psychosocial and Behavioral Health: A Systematic Review of Self-Administered Scales Validated for Postpartum Women in the United States. Womens Health Issues 25, 586-600.
Wang, L., Kroenke, K., Stump, T.E., Monahan, P.O., 2020. Screening for perinatal depression with the patient health questionnaire depression scale (PHQ-9): A systematic review and meta-analysis. General hospital psychiatry.
Wesseloo, R., Kamperman, A.M., Munk-Olsen, T., Pop, V.J., Kushner, S.A., Bergink, V., 2016. Risk of Postpartum Relapse in Bipolar Disorder and Postpartum Psychosis: A Systematic Review and Meta-Analysis. Am J Psychiatry 173, 117-127.
Yildiz PD, Ayers S, Phillips L. The prevalence of posttraumatic stress disorder in pregnancy and after birth: A systematic review and meta-analysis. J Affect Disord. 2017 Jan 15;208:634-645. doi: 10.1016/j.jad.2016.10.009.
Zubaran, C., Schumacher, M., Roxo, M., Foresti, K., 2010a. Screening tools for postpartum depression: Validity and cultural dimensions. African Journal of Psychiatry 13, 357-365.
Zubaran, C., Schumacher, M., Roxo, M.R., Foresti, K., 2010b. Screening tools for postpartum depression: Validity and cultural dimensions. African Journal of Psychiatry (South Africa) 13, 357-365.