PLOS ONE

The limited screening accuracy of the Patient Health Questionnaire-2 in detecting depression among perinatal women in Italy

Antonella Gigantesco[1]*, **Gabriella Palumbo**[1], **Loredana Cena**[2], **Laura Camoni**[1], **Alice Trainini**[2], **Alberto Stefana**[2], **Fiorino Mirabella**[1]

1 Center for Behavioural Sciences and Mental Health, National Institute of Health, Rome, Italy, **2** Department of Clinical and Experimental Sciences, Section of Neuroscience, Observatory of Perinatal Clinical Psychology, University of Brescia, Brescia, Italy

* antonella.gigantesco@iss.it

Abstract

Background

The PHQ-2 was recently recommended by the International Consortium for Health Outcomes Measurement as a form of initial perinatal screening, followed by the EPDS only for women with positive PHQ-2 score. However, the accuracy of the PHQ-2 in perinatal clinical practice has been barely researched, to date. In the present study, we aim to assess the accuracy of the PHQ-2 against the EPDS in a large sample of perinatal women.

Methods

A total of 1155 consecutive women attending eleven primary or secondary health care centres throughout Italy completed the EPDS and the PHQ-2 during pregnancy (27-40-weeks) or post-partum (1-13-weeks). Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio, negative likelihood ratio, post-test probabilities and area under the curve (AUC) of the PHQ-2, using a cut-off of ≥ 3, were calculated.

Main findings

During pregnancy, PHQ-2 revealed low sensitivity (39.5%) and PPV (39.4%) but high specificity and NPV (97.5%). In postpartum, it revealed very low sensitivity (32.7%) and moderately high NPV (80.9%), but high specificity (99.3%) and PPV (94.4%). Given the low sensitivity despite the high specificity, the PHQ-2 demonstrated poor accuracy (AUC from 0.66 to 0.68).

Conclusion

Initial screening by means of PHQ-2 failed to identify an acceptable number of perinatal women at-risk of depression in Italian clinical practice. The PHQ-2 performance suggested that it has insufficient sensitivity and discriminatory power, and may be inadequate as a screening tool for maternal depression.

Introduction

In Italy, the proportion of women who experience depression during pregnancy and postpartum ranges from 6% [1] to 12% [2, 3] and from 7 to 20% [1, 4], respectively. Perinatal depression has been associated with poorer outcomes such as a reduction in a woman's ability to perform daily activities and parenting [5], which increases the risk of significant adverse consequences over the years for the psychological well-being and health of the family [6]. It should be highlighted that maternal depression has a negative impact on the child's development [7].

Given that effective psychosocial interventions (e.g., the WHO Thinking Healthy Programme) [8] and psychological therapies (e.g., cognitive behavioural therapy and interpersonal therapy) are available [9], the importance of early detection of perinatal depression cannot be overstated. Primary and non-psychiatric specialty care visits represent important opportunities to detect or launch treatment of perinatal depression because, typically, perinatal depressed women are more likely to seek care in obstetrics/gynaecology medical settings than in specialised mental health settings. In light of the points mentioned above, several authors have called for identification procedures for perinatal depression to be introduced in clinical practice such as in obstetrics/gynaecology, paediatrics, or internal/family medicine settings [10, 11].

This demands guidelines regarding the implementation of a standardised approach to detect depression with satisfactory accuracy. It has been observed that the accuracy of depression recognition in non-psychiatric health care settings is unsatisfactory [12] and, to date, there is considerable variability regarding how the ascertainment of the risk of depression should be conducted among perinatal women in non-psychiatric clinical practice [13].

International recommendations have adopted various positions concerning the implementation of depression screening in clinical practice. In spite of the fact that the Australian Clinical Practice Guidelines for perinatal depression [14] declare that the use of a universal screening approach is good practice, and also other countries such as the United States [15] and Canada [16] recommend universal screening using the Edinburgh Postnatal Depression Scale (EPDS) [17], neither the Scottish [18], nor British [19] guidelines recommend its use. However, they do recommend selective sequential testing through a 2-step approach in which a 2-question ultra-brief instrument (i.e., Whooley questions) [20] is administered first, followed, only for some women, by the Patient Health Questionnaire-9 (PHQ-9) [21] or EPDS [17]. It should be noted, however, that the NICE recommendation to use the Whooley questions [19] was made in the absence of any validation studies in a large perinatal population. In fact, there is limited evidence for the use of the Whooley questions as a screening tool for maternal depression, only having been validated in small perinatal populations with a wide range of sensitivities found, between 46 and 100% [22].

Consistently with the recommendation of using a universal screening approach, recently, the International Consortium for Health Outcomes Measurement (ICHOM) identified the 2-item PHQ-2 as an initial step to recognize *at-risk* women to be administered at first contact during pregnancy and again during the early postpartum period, followed by the longer EPDS as confirmatory screening only for women who respond positively to either question [23].

In order to achieve a maximum likelihood that all women at risk of depressive symptoms would be administered the EPDS, it is essential that the PHQ-2 is highly sensitive and, more generally, highly accurate. However, to date, perinatal evidence of PHQ-2 accuracy is limited, and only few studies have validated the PHQ-2 in perinatal settings [22, 24]. Of these, only two have compared the PHQ-2 to a structured diagnostic interview, one showing low sensitivity and moderate specificity [25] and the other moderately high sensitivity and specificity [26]. Apart from these studies, some other studies have been conducted using the EPDS as a

reference standard, with mixed results, since sensitivity ranged from 19% to 100% and specificity ranged from 75% to 93% [27–30].

Given the paucity and the inconsistency of reports, additional research in larger and different maternity populations has been recommended in order to validate the PHQ-2 as part of a maternal health care policy to detect perinatal depression [22, 24, 29].

In the present study, we aimed to test the accuracy of the PHQ-2, using the EPDS as a reference standard, in a large sample of pregnant and postpartum women attending several primary or obstetric-gynaecology secondary care centres throughout Italy. To our knowledge, no study has yet compared the use of the PHQ-2 to the EPDS in Italian perinatal clinical practice. The main goal was to determine whether or not the PHQ-2 was highly sensitive and specific, and therefore able to rule out and rule in depression in maternity clinical practice.

Methods

Outline of the study

The study is part of a larger body of work conducted by the Observatory of Perinatal Clinical Psychology (University of Brescia, Italy) and the Italian National Institute of Health. This work included a study which merges a cross-sectional study and a pre–post intervention cohort study [31] with two main objectives: (1) to evaluate the prevalence of both maternal antepartum and postpartum depression and anxiety in a sample of women in Italy (cross-sectional study component) and (2) to evaluate the effectiveness of psychological intervention [32] for both antenatal and postnatal depression (pre-post intervention cohort study component). The present study came from the data collected in the cross-sectional study.

Study protocol

Participants were recruited from eleven publicly-funded primary or obstetric-gynaecology or paediatrics secondary care centres of the Observatory of Perinatal Clinical Psychology (University of Brescia) throughout Italy (Bergamo, Brescia, Mantua and Milan, Lombardy Region; Bologna, Emilia Romagna Region; Florence, Tuscany Region, Novara and Collegno, Piedmont Region; Roma, Lazio Region; Enna, Sicily Region). Evaluation tools were administered once during the pre- or postpartum period, depending on the characteristics of each healthcare centre. The study was approved by the ethics committee of the Healthcare Centre of Bologna (registration number 0077805, dated 6/27/2017) [29].

Procedure and participants

Participants were recruited during a routine perinatal health check-up or paediatric vaccination appointment at one of the eleven publicly-funded healthcare facilities between 2017–2018. Consecutive women were invited to participate in the study. Specifically, the participation to the study was offered by the obstetricians or gynaecologists or paediatricians of those facilities. All the women approached were provided with a pamphlet developed as part of the study, in which the purpose, aims and methodology of the study were explained. Women who wished to participate provided their personal information (name or phone number to be contacted later) in order to meet up with trained psychologists. These psychologists had attended a training on screening, assessment, and treatment for maternal perinatal mental health problems, developed by the National Institute of Health [4]. Women who definitively agreed to participate signed an informed consent form. Then, they were underwent a semi-structured interview (not a diagnostic interview) led by the trained psychologists to elicit information on current and past maternal experience with psychiatric conditions and use of psychotropic

drugs. Psychiatric conditions included symptoms of anxiety, depression, psychotic symptoms (i.e., delusions and/or hallucinations), non-suicidal self-harm tendencies, suicidal ideation or substance abuse. The inclusion criteria to be enrolled were being able to speak and read Italian well and being a woman aged ≥18 years with a biological baby aged ≤52 weeks. The exclusion criteria were having issues with drug or substance misuse and/or having on-going psychotic symptoms. The enrolled women were then administered the pertaining scheduled self-report assessment tools for data collection (see Tools). All the women completed the interviews and self-report instruments at the facilities; the majority of them on the same day in which they were invited to join. Few women provided their phone numbers to be contacted for arranging a subsequent appointment at the facilities in order to complete the instruments. The psychologists who administered the self-report tools, made sure that they returned fully completed.

Tools

Psychosocial assessment form. Socio-demographic data were collected at baseline by means of the Psychosocial Assessment Form [4] which addresses socio-demographic characteristics and other information. Socio-demographic characteristics include: age (years), marital status (married or cohabitating; single, separated, divorced or widowed), educational level (primary or illiterate; secondary high school; University), working status (student, homemaker, or unemployed; temporary employee; permanent employee), economic status (several problems; a few problems without specific difficulties; average to high status) and children living at the time of the current pregnancy/birth (yes-no).

Edinburgh Postnatal Depression Scale (EPDS). The EPDS [17, 33] is the most widely used screening tool for depressive symptoms during pregnancy and postpartum [34, 35]. It is a self-rating measure containing 10 items concerning the symptoms of depression such as anhedonia, feelings of guilt, lethargy, sleep disturbance and suicidal ideation occurring in the past 7 days. Each symptom is scored on a four-point Likert. The total score ranges from 0 to 30, with higher scores indicating more severe depressive symptoms.

Patient Health Questionnaire (PHQ-2). The PHQ-2 [36] consists of the first two items of the PHQ-9 (investigating depressed mood and anhedonia). The questions ask: "Over the last 2 weeks, how often have you been bothered by little interest or pleasure in doing things?" and "Over the last 2 weeks, how often have you been bothered by feeling down, depressed, or hopeless?". For each item, the response options are "not at all" (0), "several days" (1), "more than half the days" (2), and "nearly every day" (3). The total score ranges from 0 to 6, with higher scores indicating greater depressive symptoms.

Statistical analysis

All analyses were conducted using the Statistical Package for Social Science (SPSS) created for Windows, version 26.0. The EPDS was considered the reference standard. The PHQ-2 was analysed using total score at cut-off point 3 or more, as is usually recommended to define the result as positive [37, 38]. The EPDS total score was transformed into a binary variable to indicate positive screening for depression using the cut-off score of ≥ 13 during pregnancy, and ≥ 10 in the postpartum, as recommended by literature [39]. Socio-demographic characteristics of women with positive screening for depression were summarised using descriptive statistics. The chi-square test (or Fisher exact test) were used to test for differences between women with positive screening for depression and women without for each socio-demographic characteristic. Proportion of depression risk according to the EPDS, and positive PHQ-2 responses are presented as frequencies and percentages with 95% confidence intervals.

Internal consistency. The internal consistency of the EPDS and the PHQ-2 were assessed. Due to the fact PHQ-2 includes only two items, the mean inter-item correlation (MIC) was adopted.

Accuracy (or criterion validity). The screening accuracy of the PHQ-2, defined as sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (LR+), negative likelihood ratio (LR-), and area under the curve (AUC) against the EPDS cut-points using ROC (Receiver Operating Characteristic) analysis was assessed [40]. Positive and negative post-test probabilities were also calculated. The standard error for the area was set as non-parametric with a 95% confidence interval. Area under the curve (AUC) was interpreted as follows: AUC = 0.60–0.70 = poor, 0.70–0.80 = fair, 0.80–0.90 = good, 0.90–1.0 = excellent.

Results

A total of 1155 women, 71% of those who were asked to participate in the study, filled out both the EPDS and the PHQ-2.

Both groups of pregnant and postpartum women were primarily in their thirties (Table 1). Overall, the majority of them were married or lived with their partner and were well educated. Furthermore, the majority were employed in paid work and only few had serious economic difficulties.

Significant differences between women with positive EPDS and women without were found regarding educational level, economic problems and marital status. Specifically, among pregnant women, positive EPDS was associated with university education level and, among postnatal women, positive EPDS was associated with having several economic problems and not being married or cohabitating with a partner (Table 1).

Internal consistency

The internal consistency of the EPDS was: α = 0.80 during pregnancy and α = 0.87 following delivery. The PHQ-2 showed acceptable mean inter-item correlations (MIC) both during pregnancy (MIC = 0.38) and postpartum (MIC = 0.45).

Proportion of minor depression based on EPDS and PHQ-2 positive screen tests

The percentage of antenatal and postnatal women who screened positive for depression as determined by EPDS totalled 4.0% and 25.9%, respectively (Table 2). Positive cases determined through the PHQ-2 was 4.0% during pregnancy and 8.9% postpartum. Having a positive EPDS was found to be higher in the postpartum than during pregnancy. Regarding PHQ-2, the number of cases was also higher in the first trimester following birth.

Antepartum and postpartum minor depression: PHQ-2 accuracy

The performance of the PHQ-2 with EPDS as a standard criterion during pregnancy and postpartum is presented in Table 3.

During pregnancy, out of 38 women that scored positive on the EPDS, 15 tested positive and 23 negative using the PHQ-2. Among 38 women who screened positive using the PHQ-2, 15 were found to have probable minor antepartum depression as determined by the EPDS.

Of 916 women that scored negative with the EPDS, 893 tested negative with the PHQ-2 and among 916 women who screened negative with the PHQ-2, 893 were found not to have probable minor antepartum depression.

Table 1. Socio-demographic, pregnancy and delivery characteristics of participants. Number (valid percentage). P = statistical significance for the comparison between negative EPDS screening (score<13 during pregnancy and < 10 in postpartum) and positive EPDS screening results (score ≥ 13 during pregnancy and ≥ 10 in postpartum).

	Antenatal		Postnatal	P
	(from 27 to 40-weeks)		(from 1 to 13-weeks)	
	n (%)	EPDS ≥ 13	n (%)	EPDS ≥ 10
		n (%)		n (%)
Total	954	38 (4.0%)	201	52 (25.9%)
	(%)		(17.4%)	
Age				
18–29	209 (21.9)	8 (21.1)	30 (14.9)	6 (11.5)
30–35	453 (47.5)	19 (50.0)	79 (39.3)	16 (30.8)
> 35	292 (30.6)	11 (28.9)	92 (45.8)	30 (57.7)
Marital status				
Married or cohabitating	878 (92.6)	35 (92.1)	183 (91.5)	41 (78.8)
Single, separated, divorced or widowed	70 (7.4)	3 (7.9)	17 (8.5)	11 (21.2) [§§]
Educational level				
Primary or illiterate	100 (10.5)	9 (23.7)[§]	26 (13.0)	4 (7.7)
Secondary	341 (36.0)	11 (28.9)	83 (41.5)	21 (40.4)
University	507 (53.5)	18 (47.4)	91 (45.5)	27 (51.9)
Economic status[*]				
Several problems	58 (6.2)	5 (13.2)	16 (8.1)	9 (17.3)[§]
A few problems without specific difficulties	433 (45.9)	18 (47.4)	99 (50.0)	24 (46.2)
Average to high status	452 (47.9)	15 (39.4)	83 (41.9)	19 (36.5)
Working status				
Student, homemaker, or unemployed	151 (16.0)	7 (18.4)	39 (19.6)	8 (15.7)
Temporary employee	89 (9.4)	0 (0.0)	20 (10.1)	6 (11.8)
Permanent employee	702 (74.6)	31 (81.6)	140 (70.4)	37 (72.5)
Children living at the time of this pregnancy/birth				
No	798 (83.6)	28 (73.7)	137 (68.2)	40 (76.9)
Yes	157 (16.4)	10 (26.3)	64 (31.8)	12 (23.1)

[*]Regarding economic status: Several problems = having debts, difficulty or inability to pay daily expenses and rent; A few problems without specific difficulties = relatively modest standard of living but without particular difficulties; Average high status = home owned, possibility of taking holidays or travelling for pleasure.

[§] p<0.05.

[§§] p<0.01.

Following delivery, out of 52 women that scored positive on the EPDS, 17 tested positive and 35 negative on the PHQ-2. Among 18 women who screened positive with the PHQ-2, 17 were found to have probable minor postpartum depression.

Out of 149 women that scored negative with the EPDS, 148 tested negative with the PHQ-2 and among 183 women who screened negative with the PHQ-2, 148 were found not to have probable minor postpartum depression.

During pregnancy, the LR+ ranged from moderate (8.9) to high (27.6) and the LR- ranged from weak (0.5) to very weak (0.8) [41].

Following delivery, the LR+ ranged from moderate (6.6) to very high (357.6), with a wide and imprecise confidence interval, and the LR- was weak (from 0.56 to 0.82) [41].

The post-test probability that a woman who tested negative with the PHQ-2 had minor depression was 2.5% during pregnancy and 19.2% postpartum. The post-test probability that a

Table 2. Proportion of women with probable minor depression ascertained with EPDS, and case identification using PHQ-2.

	Pregnancy (from 27 to 40-weeks)	Postpartum (from 1 to 13-weeks)
	N = 954	N = 201
	n	n
	% (95% CI)	% (95% CI)
Screened positive for depression*	38	52
	4.0 (2.7–5.1)	25.9 (19.9–32.1)
PHQ-2˚	38	18
	4.0 (2.8–5.2)	8.9 (5.0–12.8)

* Screened positive for depression: EPDS cut-off score ≥ 13 during pregnancy and ≥ 10 postpartum.

˚ PHQ-2 total score cut-off point ≥ 3.

https://doi.org/10.1371/journal.pone.0260596.t002

woman who tested positive with the PHQ-2 had minor depression was 39.4% during pregnancy and 94.4% postpartum.

At the recommended cut-off of ≥ 3, screening accuracy of the PHQ-2 was poor (AUC = 0.66–0.68) (Table 4).

Discussion

To the best of our knowledge, this is the first study which compares the accuracy of the PHQ-2 to the EPDS in a large sample of perinatal women in Italy in order to consider its possible application as a first step screening in perinatal clinical practice.

For the PHQ-2, we used the optimal cut-off point (≥ 3) taking into consideration combined sensitivity and specificity (Youden index), as recommended by recent literature concerning the accuracy of the PHQ-2 in clinical practice [42].

Both the PHQ-2 and the EPDS demonstrated acceptable internal consistency.

In the present study, one of the main findings was that even though the benefit of including a *triage test* in perinatal clinical practice consists of narrowing the number of women who need more extensive evaluation, the PHQ-2 demonstrated that it excessively narrows the percentage of women that would need a subsequent longer assessment to a very low level of 4% of the original group of prenatal women and 8.9% of the original group of postnatal women.

In fact, the PHQ-2 demonstrated that it was poorly sensitive for identifying expectant women and new mothers at risk for depression (39.5% and 32.7%, respectively) indicating that

Table 3. Performance of the PHQ-2 against the EPDS.

	Antenatal	Postnatal
	(from 27 to 40-weeks)	(from 1 to 13-weeks)
Sensitivity (95% CI)	39.5 (24.0–56.6)	32.7 (20.3–47.1)
Specificity (95% CI)	97.5 (96.3–98.4)	99.3 (96.3–99.9)
PPV (95% CI)	39.4 (27.0–53.4)	94.4 (69.9–99.2)
NPV (95% CI)	97.5 (96.8–98.0)	80.9 (77.8–83.6)
LR+ (95% CI)	15.7 (8.9–27.6)	48.7 (6.6–357.0)
LR- (95% CI)	0.62 (0.48–0.80)	0.68 (0.56–0.82)
Positive post-test probability	39.4 (27.0–53.4)	94.4 (69.9–99.2)
Negative post-test probability	2.5 (2.0–3.2)	19.2 (16.4–22.2)

CI = Confidence Interval.

https://doi.org/10.1371/journal.pone.0260596.t003

Table 4. ROC analysis of the PHQ-2 (cut-off point ≥ 3) with EPDS as criterion standard.

Test variable	AUC	SE	Significance (p)	Lower bound	Upper bound
Screened positive for depression					
(EPDS as criterion standard)					
Pregnancy (from 27 to 40-weeks)	0.68	0.05	0.000	0.58	0.79
Postpartum (from 1 to 13 weeks)	0.66	0.049	0.001	0.56	0.76

many cases of probable depression remained undetected (60.5% and 67.3% false-negative). This sensitivity was substantially lower than that reported in the original PHQ-2 validation study (39.5% and 32.7% compared to 83%) [37] but not very distant from that reported in a recent meta-analysis which assessed the PHQ-2 pooled performance against a gold-standard diagnostic interview [24]. This meta-analysis showed that the sensitivity of the PHQ-2 in identifying major depression in primary care was lower than that reported in the original study at ≥ 3 cut-off point (pooled sensitivity 64%; with the lower boundary of the 95% CI = 46%). Consistently, a more recent systematic review has confirmed that in settings such as primary care or some inpatient and outpatients specialty care, the PHQ-2 was up to 62% sensitive for cut-off score of 3 or greater, in studies that used fully structured interviews as reference standards [43].

In contrast, we found the PHQ-2 to be highly specific (97.5%-99.3%), suggesting a low risk of false positives and response burden. As a consequence of low sensitivity, despite the high specificity, the PHQ-2 ultimately demonstrated poor accuracy, based on the ROC analyses, suggesting that it did not possess any substantial discriminatory ability both during pregnancy and postpartum (AUCs = 0.68 and 0.66, both corresponding to poor accuracy).

Looking at the PPVs, our findings seemed also to indicate a poor performance of the PHQ-2 in pregnancy, given the very low probability that a woman with a positive result indeed had depression (39.4%). In contrast, the combination of a low prevalence, low sensitivity and high specificity resulted in high NPVs (97.5% in pregnancy and 80.9% postpartum), suggesting that the probability that a woman with a negative PHQ-2 result indeed did not have depression was high, especially in pregnancy.

However, these predictive values remained uninformative, at least in prenatal screening, because they are affected by a very low prior prevalence (from 2.7% to 5.1%), which involved the base rate fallacy [44]. This makes the PHQ-2 worthless under the condition of low prevalence, which means that if you intend to use the PHQ-2, you should only apply it in pregnant populations with high prevalence of depression, such as psychiatric populations or other populations at risk of mental disorders. As evidence of this, both PPV (94.4%) and NPV (80.9%) proved to be more informative and acceptable among our sample of postnatal women, who had a higher prevalence of depression.

As pointed out by several authors [41, 45], when sensitivity is low despite high specificity, the use of the *SpPIn* mnemonics *(high **Sp**ecificity, **P**ositive, rules **In**)* is ineffective. In the present study, the *SpPIn* rule is not applicable also because of low prevalence, especially in our antenatal sample. In the same way, the use of the *SnNOut* mnemonics *(high **Sn**ensitivity, **N**egative, rules **Out**)* is not applicable because of low prevalence and low sensitivity. As a consequence, LRs and post-test probabilities may be probably the best way to evaluate the strength of the PHQ-2 in our clinical centres, where women will generally have depression at an earlier and milder stage (i.e., have a low or moderately high prevalence of depression).

The LR- of 0.62 in pregnancy and the LR- of 0.68 postpartum were weakly indicative of an absence of depression, suggesting limited accuracy in ruling out the risk of depression following a negative result [41]. After testing negative for the PHQ-2, the post-test probability of

having antenatal depression slightly decreased from 4% (pre-test probability) to 2.5%, with a confidence interval ranging from 2.0 to 3.2%. This means that a negative PHQ-2 was still compatible with a 3.2% post-test probability of depression, which clearly was low in absolute terms but unacceptably high in this situation in which the sensitivity was so low. Moreover, despite the weakly indicative LR-, the post-test probability was low because it started from a very low pre-test probability. As evidence of this, among postpartum women, with a higher pre-test probability and similar LR- of 0.68, the post-test probability resulted in a higher point estimate of 19.2% (CI: 16.4–22.2), which may be considered even more unacceptably high in this situation in which the sensitivity was even lower (32.7%). In this situation, the clinician may only feel moderately confident that depression could be ruled out. To be sure, he/she needs to advise women to proceed with other investigations.

The LR+ in the prenatal sample was 15.7 (CI: 8.9–27.6), suggesting that the PHQ-2 ranged from moderately (estimated shift in probability of at least 30%, considering the lower boundary of the CI) to strongly (estimated shift in probability of at least 45%, considering the upper boundary of the CI) indicative of the presence of depression. The post-test probability of having prenatal depression increased from 4.0% (pre-test probability) to 39.4% (CI: 27.0–53.4). Despite this considerable increase, the post-test probability revealed the presence of depression among only about one in 2–3 women with a positive PHQ-2, under the low pre-test probability and sensitivity. This may make the clinician decide to rule in prenatal depression and proceed with a further screening assessment, but with moderate confidence.

The LR+ in the postnatal sample was 48.7 with a very wide confidence interval (95% CI: 6.6–357.0). The probability of having postnatal depression increased from 25.9% to 94.4% (95% CI: 69.9–99.2) revealing that about one in 1–1.4 women with a positive test had depression. Despite the fact the LR+ suffered from a very wide confidence interval, with the lower 95% confidence interval including 6.6, which was moderately indicative of the presence of depression, this may make the clinician decide to rule in women for the presence of postnatal depression and proceed with the second step of a two-stage screening strategy, but with moderate confidence.

The few available comparable studies are mostly consistent with our findings. In particular, our findings are very consistent with those of a recent Australian study which used the same EPDS cut-offs to test the accuracy of the PHQ-2 for probable minor depression [29]. In that study, during pregnancy (36-weeks), sensitivity, specificity, PPV, NPV, LR+ and LR- were 0.50, 0.95, 0.30, 0.98, 9.4 and 0.53, respectively, and at 6-weeks postpartum were 0.39, 0.97, 0.68, 0.91, 14.59 and 0.62, respectively. In the same study, using ROC analysis, the PHQ-2 cut-off ≥ 3 was found to be poor to fair for minor depression (AUC = 0.68 in pregnancy; AUC = 0.72 postpartum), similarly to our AUC findings of 0.66–0.68. Our findings were also similar to those of a study of Smith and colleagues [25], which reported low PHQ-2 sensitivity (62%). Very consistently with the present study, another postnatal study [27] showed a sensitivity of 43.5% and a specificity of 97.2%, using the same PHQ-2 cut-off point of 3 or more to identify a positive screen with the EPDS (cut-off point of 10 or more).

A limited number of other studies have reported higher sensitivity. Gjerdingen et al., reported that the PHQ-2 had 75% sensitivity, 88% specificity, 24% PPV and 99% NPV, in a sample of postpartum women (0–1 month postpartum) [26]. However, Gjerdingen et al., used a different cut-off of 2 or higher, which maximized sensitivity. Bennett et al., [28] found that PHQ-2 had a sensitivity of 93%, 82%, and 80%, and specificity of 75%, 80%, and 86%, at 15 and 30 weeks gestational age and 6–16 weeks postpartum, respectively. However, they found low PPVs (44, 24, and 30% respectively), similarly to our antenatal PPV. Furthermore, Bennett et al., adopted a different PHQ-2 response format (dichotomous yes/no, where responding yes to either of the items was considered a positive result), which maximised sensitivity and

specificity [29] and a sample with low education levels. It should be noted that it was found that PHQ-2 sensitivity was higher for women who went on beyond high school education than for those with a high school education or less [27]. Chae et al., [30], finally, found that the sensitivity of the PHQ-2 was 100% and the specificity was 79.3% among postpartum women attending a family multi-ethnic medicine residency centre. However, also in this case, the Authors adopted a dichotomous PHQ-2 response format which maximised sensitivity and specificity.

Overall, these studies [26–28, 30] suggest that lowering the cut-off point to 2 or higher will increase sensitivity. This, however, would come at the cost of lowered specificity given its inverse relationship with sensitivity. Moreover, at this lower cut-off the specificity is likely to be further reduced when the prevalence of depression is low [24, 29]. This suggests that the PHQ-2 at a cut-off point of 2 or more may have limited usefulness in the identification of women with depression in primary or secondary care services, because in such contexts the prevalence of depression is likely to be low. The extent to which lowering the cut-off point would be a valid option depends on the prevalence of depression and the cost and availability of subsequent strategies, potentially burdensome in busy maternity settings, to further assess those who score positively on the initial screening.

Strengths and limitations

The strengths of the present study include the use of a large sample and several maternity clinical centres throughout Italy. The limitations include a low prevalence of antenatal depression and the use of the EPDS as a reference standard instead of a structured diagnostic interview. However, our prevalence is only slightly less than and comparable to those reported by other Italian studies [3, 4], and the aim in the present study was not to evaluate the diagnostic accuracy of the PHQ-2 but rather its screening accuracy compared to a longer more accredited screening instrument. In other words, the aim in the present study was to investigate if the PHQ-2 was accurate enough to be used as a substitute for a longer screening tool (as would occur in clinical practice using the ICHOM recommended procedure), not for a diagnostic instrument which formally assessed major depression.

Conclusion

This study found limited evidence regarding the use of the PHQ-2 as part of a perinatal multistage case finding strategy at the recommended cut-off point of ≥ 3 [23, 37, 42]. Other Authors have shown that similar ultra-rapid screening instruments, such as the Whooley questions [20], had very low sensitivity when routinely administered in early pregnancy [46].

Maternity and primary care services require simple, quick screening tools to know who to refer. The present study calls into question the appropriateness of the PHQ-2, which aligns with the literature raising questions about the validity of the Whooley questions. The findings in this study may impact on service provision and making decisions about which screening instruments to use. It suggests that more comprehensive instruments may be needed as a first line. In our opinion, directly administering a longer screening instrument such as the EPDS may be the best option. Should this be difficult in a busy maternity setting, a tablet or paper could represent a useful alternative [46].

Supporting information

S1 Raw data. Data set on the accuracy of the Patient Health Questionnaire-2 in detecting depression among perinatal women in Italy.
(XLSX)

Acknowledgments

Author Contributions

Conceptualization: Antonella Gigantesco.

Data curation: Antonella Gigantesco.

Formal analysis: Fiorino Mirabella.

Investigation: Loredana Cena.

Methodology: Antonella Gigantesco.

Project administration: Loredana Cena.

Supervision: Alberto Stefana.

Writing – original draft: Antonella Gigantesco.

Writing – review & editing: Antonella Gigantesco, Gabriella Palumbo, Laura Camoni, Alice Trainini, Alberto Stefana.

References

1. Cena L, Mirabella F, Palumbo G, Gigantesco A, Trainini A, Stefana A. Prevalence of maternal antenatal and postnatal depression and their association with sociodemographic and socioeconomic factors: A multicentre study in Italy. J Affect Disord. 2021; 279: 217–21. https://doi.org/10.1016/j.jad.2020.09.136 PMID: 33069119

2. Corbani IE, Rucci P, Iapichino E, Quartieri Bollani M, Cauli G, Ceruti MR, et al. Comparing the prevalence and the risk profile for antenatal depressive symptoms across cultures. The Int J Soc Psychiatry. 2017; 63(7): 622–31. https://doi.org/10.1177/0020764017725543 PMID: 28805152

3. Banti S, Mauri M, Oppo A, Borri C, Rambelli C, Ramacciotti D, et al. From the third month of pregnancy to 1 year postpartum. Prevalence, incidence, recurrence, and new onset of depression. Results from the perinatal depression-research & screening unit study. Compr Psychiatry. 2011; 52(4):343–51. https://doi.org/10.1016/j.comppsych.2010.08.003 PMID: 21683171

4. Palumbo G, Mirabella F, Gigantesco A. Positive screening and risk factors for postpartum depression. Eur Psychiatry. 2017; 42:77–85. https://doi.org/10.1016/j.eurpsy.2016.11.009 PMID: 28212509

5. Glover V. Maternal depression, anxiety and stress during pregnancy and child outcome; what needs to be done. Best Pract Res Clin Obstet Gynaecol. 2014; 28(1):25–35. https://doi.org/10.1016/j.bpobgyn.2013.08.017 PMID: 24090740

6. Caspi A, Houts RM, Belsky DW, Harrington H, Hogan S, Ramrakha S, et al. Childhood forecasting of a small segment of the population with large economic burden. Nat Hum Behav. 2016; 1:0005. https://doi.org/10.1038/s41562-016-0005 PMID: 28706997

7. Milgrom J, Schembri C, Ericksen J, Ross J, Gemmill AW. Towards parenthood: an antenatal intervention to reduce depression, anxiety and parenting difficulties. J Affect Disord. 2011; 130(3):385–94. https://doi.org/10.1016/j.jad.2010.10.045 PMID: 21112641

8. Rahman A, Malik A, Sikander S, Roberts C, Creed F. Cognitive behaviour therapy-based intervention by community health-workers for mothers with depression and their infants in rural Pakistan: cluster-randomised controlled trial. Lancet. 2008 Sep 13; 372(9642): 902–9. https://doi.org/10.1016/S0140-6736(08)61400-2 PMID: 18790313

9. Howard LM, Khalifeh H. Perinatal mental health: a review of progress and challenges. World Psychiatry. 2020 Oct; 19(3):313–27. https://doi.org/10.1002/wps.20769 PMID: 32931106

10. Legere LE, Wallace K, Bowen A, McQueen K, Montgomery P, Evans M. Approaches to health-care provider education and professional development in perinatal depression: a systematic review. BMC Pregnancy Childbirth. 2017; 17(1):239. https://doi.org/10.1186/s12884-017-1431-4 PMID: 28738855

11. Sit DK, Wisner KL. Identification of postpartum depression. Clin Obstet Gynecol. 2009; 52(3):456–68. https://doi.org/10.1097/GRF.0b013e3181b5a57c PMID: 19661761

12. Mitchell AJ. Why do clinicians have difficulty detecting depression? In: Mitchell AJ, Coyne JC, editors. Screening for depression in clinical practice. New York: Oxford University Press; 2010. pp. 57–82.

13. Austin MP, Marcé Society Position Statement Advisory Committee. Marcé International Society position statement on psychosocial assessment and depression screening in perinatal women. Best Pract Res Clin Obstet Gynaecol. 2014; 28(1):179–87. https://doi.org/10.1016/j.bpobgyn.2013.08.016 PMID: 24138943

14. Austin MP, Highet N and the Guidelines Expert Advisory Committee. Clinical practice guidelines for depression and related disorders–anxiety, bipolar disorder and puerperal psychosis–in the perinatal period. A guideline for primary care health professionals. Melbourne, beyondblue: the national depression initiative. Melbourne, Australia 2011.

15. American College of Obstetricians and Gynaecologists. ACOG Committee Opinion No. 757: screening for perinatal depression. Obstet Gynecol. 2018; 132(5):e208 ee212. https://doi.org/10.1097/AOG.0000000000002927 PMID: 30629567

16. BC Reproductive Mental Health Program and Perinatal Services BC. 2014 (accessed 18 August 2019). Best practice guidelines for mental health disorders in the perinatal period. Available from: http://www.perinatalservicesbc.ca/Documents/Guidelines-Standards/ Maternal/MentalHealthDisordersGuideline.pdf.

17. Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression: development of the 10-item Edinburgh postnatal depression scale. Br J Psychiatry. 1987; 150:782–6. https://doi.org/10.1192/bjp.150.6.782 PMID: 3651732

18. Scottish intercollegiate guidelines network (SIGN). 2012 March (accessed 7 July 2016). Management of perinatal mood disorders (SIGN Publication no. 127). Available from: http://www.sign.ac.uk.

19. National Institute for health and Care Excellence (NICE). February 2007 (accessed 7 July 2016). Antenatal and post- natal mental health: clinical management and service guidance. NICE Clinical Guideline CG45. Available from: http://www.nice.org.uk/guidance/cg45.

20. Whooley MA, Avins AL, Miranda J, Browner WS. Case-finding instruments for depression. Two questions are as good as many. J Gen Intern Med 1997; 12 (7):439–45. https://doi.org/10.1046/j.1525-1497.1997.00076.x PMID: 9229283

21. Kroenke K, Spitzer RL, and Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001; 16 (9):606–13. https://doi.org/10.1046/j.1525-1497.2001.016009606.x PMID: 11556941

22. Mann R, Adamson J, Gilbody SM. Diagnostic accuracy of case-finding questions to identify perinatal depression. CMAJ 2012; 184(8): E424–30. https://doi.org/10.1503/cmaj.111213 PMID: 22451686

23. International Consortium for Health Outcomes Measurement (ICHOM). 2019 (accessed 9th April 2019). Pregnancy & childbirth. Data collection reference guide. Version 1.1. Available from: https://ichom.org/files/medical-conditions/pregnancy-and-childbirth/ pregnancychildbirth-reference-guide.pdf

24. Manea L, Gilbody S, Hewitt C, North A, Plummer F, Richardson R, et al. Identifying depression with the PHQ-2: A diagnostic meta-analysis. J Affect Disord. 2016; 203:382–95. https://doi.org/10.1016/j.jad.2016.06.003 PMID: 27371907

25. Smith MV, Gotman N, Lin H, Yonkers KA. Do the PHQ-8 and the PHQ-2 accurately screen for depressive disorders in a sample of pregnant women? Gen Hosp Psychiatry. 2010; 32(5):544–8. https://doi.org/10.1016/j.genhosppsych.2010.04.011 PMID: 20851275

26. Gjerdingen D, Crow S, McGovern P, Miner M, Center B. Postpartum depression screening at well-child visits: validity of a 2-question screen and the PHQ-9. Ann Fam Med. 2009; 7(1):63–70. https://doi.org/10.1370/afm.933 PMID: 19139451

27. Cutler CB, Legano LA, Dreyer BP, Fierman AH, Berkule SB, Lusskin SI, et al. Screening for maternal depression in a low education population using a two item questionnaire. Arch Women's Ment Health. 2007; 10(6):277–83. https://doi.org/10.1007/s00737-007-0202-z PMID: 17710366

28. Bennett IM, Coco A, Coyne JC, Mitchell AJ, Nicholson J, Johnson E et al. Efficiency of a two-item pre-screen to reduce the burden of depression screening in pregnancy and postpartum: an IMPLICIT network study. J Am Board Fam Med. 2008; 21(4):317–25. https://doi.org/10.3122/jabfm.2008.04.080048 PMID: 18612058

29. Slavin V, Creedy DK, Gamble J. Comparison of screening accuracy of the Patient Health Questionnaire-2 using two case-identification methods during pregnancy and postpartum. BMC Pregnancy Childbirth. 2020; 20(1):211. https://doi.org/10.1186/s12884-020-02891-2 PMID: 32290813

30. Chae SY, Chae MH, Tyndall A, Ramirez MR, Winter RO. Can we effectively use the two-item PHQ-2 to screen for postpartum depression? Fam Med. 2012 Nov-Dec; 44(10):698–703. PMID: 23148001

31. Cena L, Palumbo G, Mirabella F, Gigantesco A, Stefana A, Trainini A et al. Perspectives on early screening and prompt intervention to identify and treat maternal perinatal mental health. Protocol for a

prospective multicenter study in Italy. Front Psychol. 2020; 11:365. https://doi.org/10.3389/fpsyg.2020.00365 PMID: 32218756

32. Milgrom J, Martin PR, and Negri LM. Treating Postnatal Depression: A Psychological Approach for Health Care Practitioners. Chichester: Wiley; 1999.

33. Benvenuti P, Ferrara M, Niccolai C, Valoriani V, Cox JL. The Edinburgh postnatal depression scale: validation for an Italian sample. J Affect Disord. 1999; 53(2):137–41. https://doi.org/10.1016/s0165-0327(98)00102-5 PMID: 10360408

34. de Moraes GPA, Lorenzo L, Pontes GAR, Montenegro MC, Cantilino A. Screening and diagnosing postpartum depression: when and how? Trends Psychiatry Psychother. 2017; 39(1): 54–61. https://doi.org/10.1590/2237-6089-2016-0034 PMID: 28403324

35. Ukatu N, Clare CA, Brulja M. Postpartum depression screening tools: a review. Psychosomatics 2018; 59(3): 211–9. https://doi.org/10.1016/j.psym.2017.11.005 PMID: 29396166

36. Löwe B, Kroenke K, Gräfe KJ. Detecting and monitoring depression with a two-item questionnaire (PHQ-2). J Psychosom Res. 2005; 58(2):163–71. https://doi.org/10.1016/j.jpsychores.2004.09.006 PMID: 15820844

37. Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: validity of a two-item depression screener. Med Care. 2003; 41(11):1284–92. https://doi.org/10.1097/01.MLR.0000093487.78664.3C PMID: 14583691

38. Nijagal MA, Wissig S, Stowell C, Olson E, Amer-Wahlin I, Bonsel G, et al. Standardized outcome measures for pregnancy and childbirth, an ICHOM proposal. BMC Health Serv Res. 2018; 18(1):953. https://doi.org/10.1186/s12913-018-3732-3 PMID: 30537958

39. Matthey S, Henshaw C, Elliott S, Barnett B. Variability in use of cut-off scores and formats on the Edinburgh postnatal depression scale–implications for clinical and research practice. Arch Women's Ment Health. 2006; 9(6): 309–15. https://doi.org/10.1007/s00737-006-0152-x PMID: 17013761

40. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978; 8(4):283–98. https://doi.org/10.1016/s0001-2998(78)80014-2 PMID: 112681

41. Baeyens JP, Serrien B, Goossens M, Clijsen R. Questioning the "SPIN and SNOUT" rule in clinical testing. Arch Physiother. 2019; 9:4. https://doi.org/10.1186/s40945-019-0056-5 PMID: 30891312

42. Mitchell AJ, Yadegarfar M, Gill J, Stubbs B. Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies. BJPsych Open. 2016; 2(2):127–38. https://doi.org/10.1192/bjpo.bp.115.001685 PMID: 27703765

43. Levis B, Sunm Y, Hem C, Wum Y, Krishnanm A, Bhandarim PM, et al. Accuracy of the PHQ-2 Alone and in Combination With the PHQ-9 for Screening to Detect Major Depression: Systematic Review and Meta-analysis. JAMA. 2020 Jun 9; 323(22):2290–2300. https://doi.org/10.1001/jama.2020.6504 PMID: 32515813

44. Barhillel M. The base-rate fallacy in probability judgments. Acta Psychologica 1980; 44(3):211–33. https://doi.org/10.1016/0001-6918(80)90046-3

45. Pewsner D, Battaglia M, Minder C, Marx A, Bucher HC, Egger M. Ruling a diagnosis in or out with "SpPIn" and "SnNOut": a note of caution. BMJ. 2004; 329(7459):209–13. https://doi.org/10.1136/bmj.329.7459.209 PMID: 15271832

46. Howard LM, Ryan EG, Trevillion K, Anderson F, Bick D, Bye A, et al. Accuracy of the Whooley questions and the Edinburgh Postnatal Depression Scale in identifying depression and other mental disorders in early pregnancy. Br J Psychiatry. 2018; 212(1):50–6. https://doi.org/10.1192/bjp.2017.9 PMID: 29433610