Psychiatry Research 332 (2024) 115687

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Optimal risk and diagnosis assessment strategies in perinatal depression: A machine learning approach from the life-ON study cohort

Armando D'Agostino [a,b,*], Corrado Garbazza [c,d,e], Daniele Malpetti [f], Laura Azzimonti [f], Francesca Mangili [f], Hans-Christian Stein [a], Renata del Giudice [b], Alessandro Cicolin [g], Fabio Cirignotta [h], Mauro Manconi [e,i,j]

[a] Department of Health Sciences, Università degli Studi di Milano, Italy
[b] Department of Mental Health and Addiction, ASST Santi Paolo e Carlo, Milan, Italy
[c] Centre for Chronobiology, University of Basel, Basel, Switzerland
[d] Transfaculty Research Platform Molecular and Cognitive Neurosciences, University of Basel, Basel, Switzerland
[e] Sleep Medicine Unit, Neurocenter of Southern Switzerland, Lugano, Switzerland
[f] Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), USI/SUPSI, Lugano, Switzerland
[g] Department of Neuroscience, Sleep Medicine Center, University of Turin, Turin, Italy
[h] University of Bologna, Italy
[i] Faculty of Biomedical Sciences, Università della Svizzera Italiana, Lugano, Switzerland
[j] Department of Neurology, University Hospital, Inselspital, Bern, Switzerland

ARTICLE INFO

Keywords:
Postpartum depression
Depression risk prediction
Major depressive episode (MDE)
Montgomery–Åsberg depression rating scale (MADRS)
Hamilton depression rating scale (HDRS)
Edinburgh postnatal depression scale (EPDS)
Visual analog scale (VAS)

ABSTRACT

This study aimed to assess the concordance of various psychometric scales in detecting Perinatal Depression (PND) risk and diagnosis. A cohort of 432 women was assessed at 10–15th and 23–25th gestational weeks, 33–40 days and 180–195 days after delivery using the Edinburgh Postnatal Depression Scale (EPDS), Visual Analogue Scale (VAS), Hamilton Depression Rating Scale (HDRS), Montgomery-Åsberg Depression Rating Scale (MADRS), and Mini International Neuropsychiatric Interview (MINI). Spearman's rank correlation coefficient was used to assess agreement across instruments, and multivariable classification models were developed to predict the values of a binary scale using the other scales. Moderate agreement was shown between the EPDS and VAS and between the HDRS and MADRS throughout the perinatal period. However, agreement between the EPDS and HDRS decreased postpartum. A well-performing model for the estimation of current depression risk (EPDS $>$ 9) was obtained with the VAS and MADRS, and a less robust one for the estimation of current major depressive episode (MDE) diagnosis (MINI) with the VAS and HDRS. When the EPDS is not feasible, the VAS may be used for rapid and comprehensive postpartum screening with reliability. However, a thorough structured interview or clinical examination remains necessary to diagnose a MDE.

1. Introduction

Perinatal Depression (PND) is generally considered a Major Depressive Episode (MDE) occurring at any time during pregnancy and up to 12 months after delivery (ACOG, 2018). Although adjusted pooled prevalence estimates reach 12% of all pregnancies, considerable variability has been reported by studies using symptom scales or diagnostic instruments (Woody et al., 2017). A nearly detection of PND is important but the identification of depressive episodes in the perinatal period poses several clinical challenges. The term depression encompasses a broad range of symptoms that might at times be difficult to distinguish from physiological distress reactions (Snaith, 1996). Although somewhat time-consuming, many studies have shown structured interviews to improve performance in case classification and prevalence estimates of disorders (Mitchell et al., 2011; Moussavi et al., 2007). Nonetheless, a recent umbrella review of 69 meta-analyzes including 81 prevalence estimates found that only 10% reflected studies that classified depression exclusively through clinician–administered structured interviews, whereas almost 90% used screening or rating tools alone, or in combination with other methods (e.g., medical records, self-report); pooled prevalence rates varied considerably, ranging from 17% for diagnostic interviews, to 22% and 31% for studies based on combinations and

* Corresponding author at: Department of Health Sciences, Università degli Studi di Milano, Via Antonio di Rudinì, 8, Milano 20142, Italy.
E-mail address: armando.dagostino@unimi.it (A. D'Agostino).

https://doi.org/10.1016/j.psychres.2023.115687
Received 17 July 2023; Received in revised form 15 December 2023; Accepted 20 December 2023
Available online 24 December 2023

screening or rating tools, respectively (Levis et al., 2019b).

Given the abundance of instruments used for PND, direct comparisons in large samples are necessary to select optimal, disorder-specific tools. The self-administered Edinburgh Postnatal Depression Scale (EPDS) is the most widely used screening tool for PND in community-based settings (Bhat et al., 2022). Above the identified cut-off score, women should be referred to a specialist to verify diagnosis with a structured clinical interview - such as the Mini International Neuropsychiatric Interview (MINI) (Lecrubier et al., 1997; Sheehan et al., 1997) or the Structured Clinical Interview for DSM disorders (First et al., 2016). The MINI has recently been found to classify more patients with depression than other instruments by an individual participant data meta-analysis of 57 studies including 17,158 participants (Levis et al., 2018).

The Hamilton Depression Rating Scale (HDRS; Hamilton 1960) and the Montgomery-Asberg Depression Rating Scale (MADRS; Montgomery and Asberg 1979) are commonly employed clinician-administered rating scales to measure treatment response in clinical trials (Cuijpers et al., 2021; Sockol et al., 2011), to assess changes in symptom severity over time in clinical practice and as relatively brief screening tools for depression. Unlike the HDRS, the MADRS mainly targets core mood symptoms, such as sadness, tension, pessimistic thoughts, and suicidal thoughts.

Recent developments in machine learning (ML) models for predicting postpartum depression risk involve combining clinical, sociodemographic, and biological data. A predominant focus on supervised learning and common ML models like support vector machines, random forest, and logistic regression has been observed (Zhong et al., 2022). Notably, patient psychiatric and gynecological history, along with sociodemographic information, have proven reliable in identifying those at risk (Cellini et al., 2022; Wakefield and Frasch, 2023; Xu and Sampson, 2023). Although fewer studies have explored biological variables, differences in metabolite changes show promise in classifying women with and without postpartum depression (Yu et al., 2022). Recent ML applications to postpartum depression risk include a study utilizing patient-reported survey responses in early pregnancy, demonstrating moderate performance in predicting depression risk across trimesters and postpartum periods (Reps et al., 2022). Another study identified mood status in the first trimester, previous depressive episodes, and marital status as crucial predictors of later onset postpartum depression, with additional factors such as sleep quality, age, previous miscarriages, and adverse life events enhancing predictive model performance (Garbazza et al., under review).

In this study, we aimed to investigate the relationship between a structured clinical interview (MINI) and both clinician-administered and self-report validated scales in a large cohort of women assessed at multiple time points from the first trimester of pregnancy to the first six months postpartum. Our objectives were to compare classification models for perinatal depression risk and diagnosis using a multiscale assessment approach, establish consistency across instruments, and determine an optimal selection of assessment tools both before and after delivery.

2. Materials and methods

2.1. Data collection

Data for the present analysis were extracted from the "Life-ON" study, a multicenter, prospective cohort study on sleep and mood changes in the perinatal period, which has been extensively described elsewhere (Baiardi et al., 2016; Garbazza et al., 2022). The recruitment of participants in the 4 centers was carried on for 3 years, between the beginning of 2016 and 2019. The last follow-up visit, falling 18 months after the inclusion of the last participant, took place in June 2020. In total, about 2000 women in the first trimester of pregnancy were contacted and invited to participate in the study.

Four hundred and thirty-nine women (age: mean 33.7, std 4.2) were longitudinally followed-up from the first trimester of pregnancy until 12 months postpartum. Main inclusion criteria were age 18–45 years, gestational age between 10 and 15 weeks, lack of major medical conditions; main exclusion criteria were a diagnosis of bipolar disorder or psychosis, a current or recent (within 6 months) depressive episode (Baiardi et al., 2016). Five different rating scales for depression were administered at four different time points during the study: visits 1 (10–15th gestational week), 3 (23–25th gestational week), 6 (33–40 days after delivery) and 9 (180–195 days after delivery). From these data, complete combinations of scores on all five depression scales were obtained. When all 5 scales were administered to the participants, their completion required up to 1 h of time. However, this occurred in 5 out of 11 study visits, while in the remaining 6 follow-up visits the only psychiatric scales administered were the EPDS and VAS, which required an average of 5 min to be completed. All clinician–administered scales and the MINI structured interviews were conducted by staff psychologists or physicians. Participants were first asked to complete the self-administered scales EPDS and VAS by themselves. This was to respect a consistent time sequence, given that in 6 visits EPDS and VAS were the only two scales administered. Furthermore, in this way we tried to avoid that the participants could be influenced in the self-assessment of their mood by the subsequent clinical interview with the investigator. Afterwards, when required according to the study protocol, the investigator carried out a clinical interview based on the MINI scale. Finally, the semi-structured HDRS and MADRS scales were administered by the researcher and discussed together with the participants. Multi-scale combinations were available for 432 different women, and for 421, 336, 277 and 242 women at visit 1, 3, 6 and 9, respectively. For 195 women, combinations of scores were available at all four visits. In terms of prepartum and postpartum, 757 observations came from the prepartum phase (visits 1 and 3) and 519 observations from the postpartum phase (visits 6 and 9). A total of 1276 complete score combinations were extracted from the Life-ON cohort, to constitute the pooled data set. Fig. 1 summarizes the distribution and dispersion of the sample throughout the study visits.

The five scales administered during the study were:

- The Hamilton Rating Scale for Depression (HDRS) has two common versions with either 17 or 21 items and is scored between 0 and 4 points. The first 17 items measure the severity of depressive symptoms while the extra four items on the extended 21-point scale measure factors that might be related to depression, but are not thought to be measures of severity, such as paranoia or obsessional and compulsive symptoms. Here we used the 17-item scale, which yields the following total score ranges: 0–7 (no depression), 8–16 (mild depression), 17–23 (moderate depression), >24 (severe depression).
- The Montgomery-Äsberg Depression Rating Scale (MADRS), consists of 10 items evaluating core symptoms of depression. Nine of the items are based upon patient reports, and one is on the rater's observation during the rating interview. MADRS items are rated on a 0–6 continuum (0 = no abnormality, 6 = severe), yielding the following total score ranges: 0–6 (no depression), 7–19 (mild depression), 20–34 (moderate depression), >34 (severe depression).
- The Edinburgh Postnatal Depression Scale (EPDS) is a 10-item self-administered screening tool used tailored for women during pregnancy and postpartum. Responses are scored 0–3 according to the severity of the symptom. The total score is determined by adding together the scores for each of the 10 items. The scale is considered an effective screening tool for major and minor depressive syndromes throughout pregnancy and postpartum above a total score of 9 (Levis et al., 2019a), but its accuracy increases if the cut-off is raised above 12 (Cox, 2019). In this study, we chose the lower cut-off to favor a more inclusive approach because the study cohort was composed of women without a diagnosis of depression at baseline.

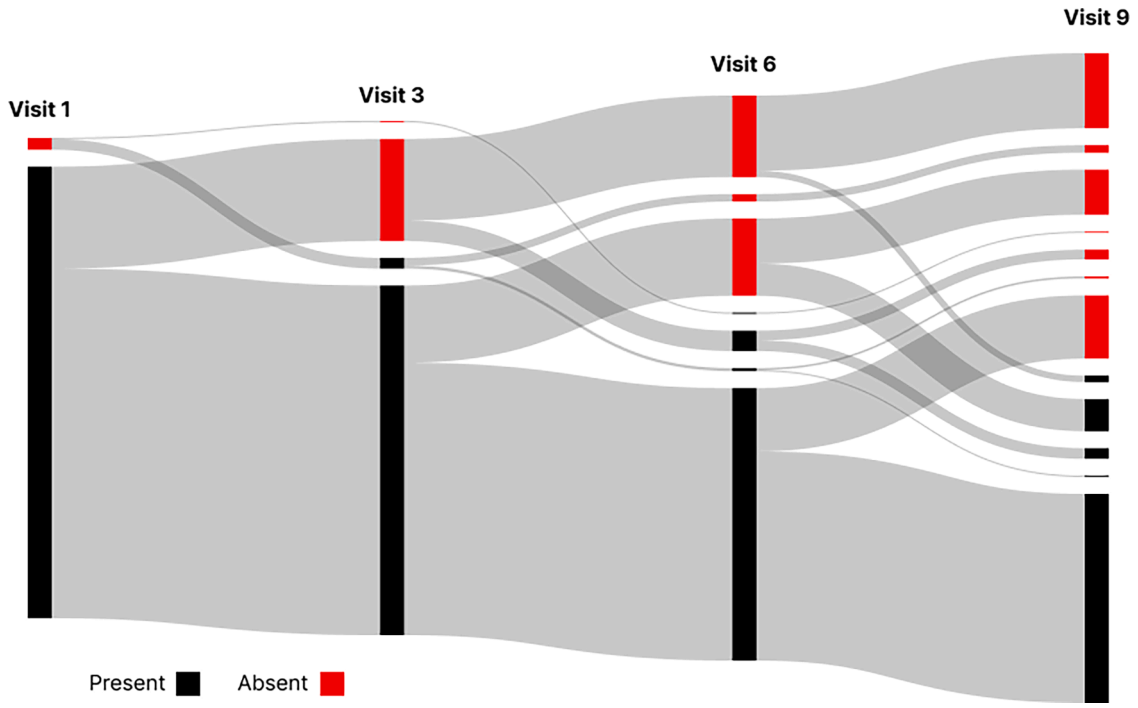

Fig. 1. Alluvial plot displaying the progression of participation throughout the study. Each vertical block's height represents the number of patients for a particular visit, with black blocks indicating patients attending and red blocks representing those who missed the visit. The blocks are interconnected to demonstrate how they change over time.

Moreover, in Italian validation studies optimal cut-offs were found to be 9/10 (Carpiniello et al., 1997), or 8/9 in the context of community screenings (Benvenuti et al., 1999).

- The self-reported single-item visual analogue scale (VAS) is used to evaluate depression severity with the following instruction: "On a scale from 0 to 10, where 0 is the worst mood imaginable and 10 is the best mood imaginable, please indicate how you are feeling right now by marking a point on the line". Participants rate their self-perceived level of depression by making a cross on a continuous, straight 10 cm-line drawn on paper. Outcome values are calculated by measuring the distance reached from point 0 using a ruler.
- The Mini International Neuropsychiatric Interview (MINI), is a short, fully structured interview designed to identify the 17 most common psychiatric disorders in DSM-III-R, DSM-IV, DSM-5 and ICD-10.

For the aims of this study, we considered non-binary, continuous scores for HDRS, MADRS, EPDS and VAS, and binary scores for the MINI (presence or absence of a depressive episode), and the EPDS (above or below the cut-off).

2.2. Statistics

The Mann-Whitney U test was applied to different non-binary scales to assess differences in distribution between the following samples: (i) prepartum and postpartum samples, (ii) samples with positive and negative MINI, (iii) samples with EPDS above 9 or not. In all cases, since we performed multiple tests, we adjusted the p-values for multiple tests by means of the Benjamini-Hochberg correction (Benjamini and Hochberg, 2019), to control the false discovery rate.

We regarded the presence of a monotonic relationship as indicative of agreement between two non-binary scales, and we assessed this by means of Spearman's rank correlation coefficient. All the analyzes were performed in R (Core Team, 2018).

2.3. Classification models

We considered the full available sample (including both pregnancy and the postpartum), and focused on the prediction of a binary scale using the values of all the other available scales for the same patient at the same visit. In particular, we focused on predicting MINI (using HDRS, MADRS, EPDS and VAS), and a binarized form of EPDS (using HDRS, MADRS, VAS and MINI). This latter was defined by considering women with EPDS$>$9 as at-risk for depression ("Yes") and women with EPDS$\leq$9 as not at-risk for depression ("No").

We developed multivariable classification models for predicting the values of a binary scale using the other scales. The performances were evaluated in cross validation (five-fold) with 10 repetitions, to assess the stability of the obtained results. Stratification with respect to both target and visit values was used for creating folds. Moreover, the division into folds for the n-th repetition of any two models predicting the same target was chosen to be the same. The optimal probability thresholds for assigning outcome classes to model predictions was selected by optimizing the geometric mean of sensitivity and specificity in a nested cross validation set up. The classification models were implemented by means of the caret R package (Kuhn, 2021).

As evaluation metrics for the performances of the classification models, we used the area under receiving operating curve (AUROC), the area under the precision recall curve (AUPR), sensitivity (SEN) and specificity (SPE). We pooled together predictions for the different folds of a same repetition, calculated the values of the metrics for the single repetitions, and then calculated mean and standard deviation across the ten repetitions.

3. Results

3.1. Overall distributions of rating scale scores during pregnancy and the postpartum

We analyzed the overall distributions of the five scales, shown as histograms in Fig. 2A. As highlighted by the plots, there is a large

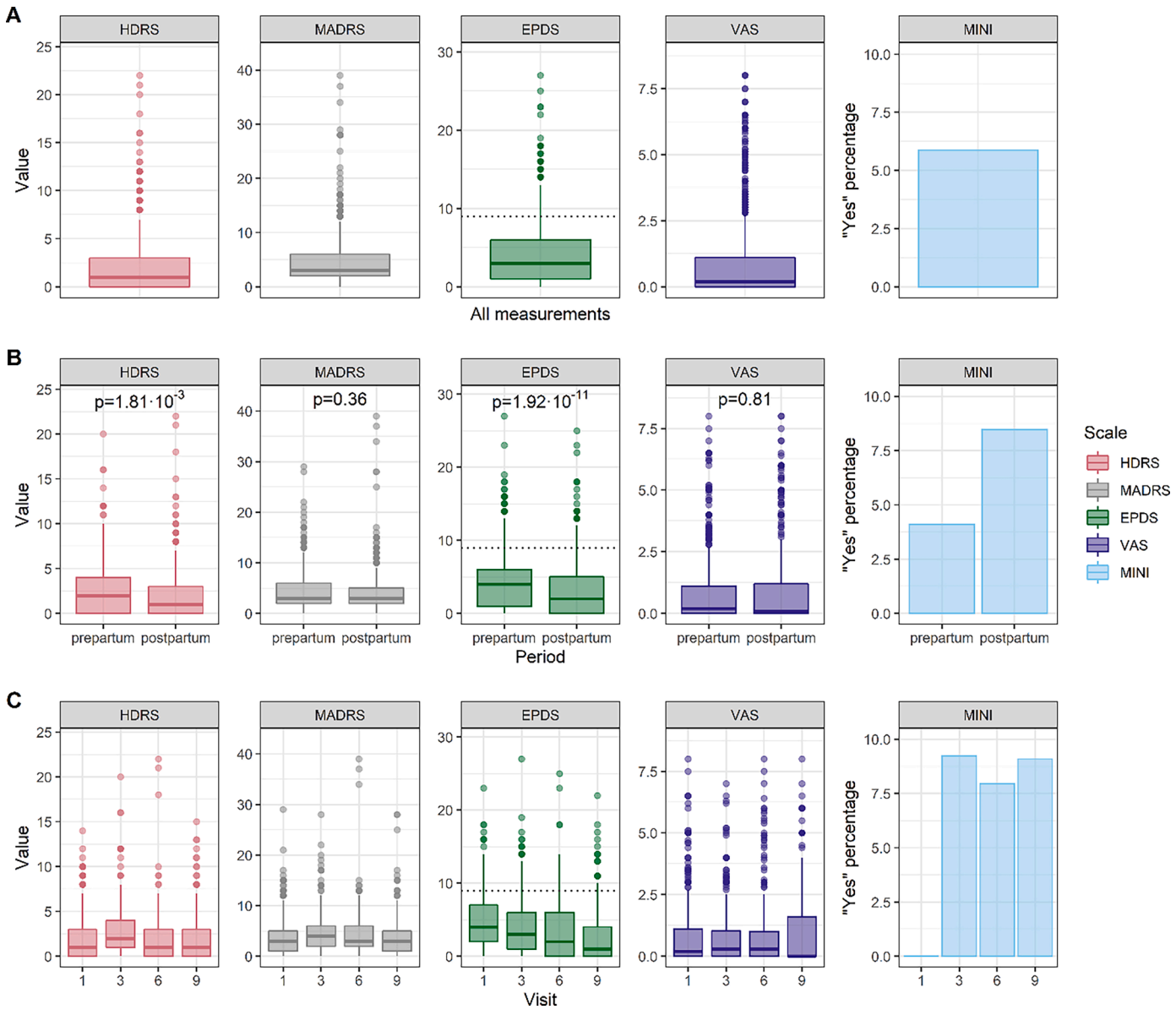

Fig. 2. Distribution of measurements for the different scales. In each subfigure, for non-binary scales, boxplots show the distribution of values; for MINI, which has the two possible values "Yes" and "No", a bar plot shows the percentage of "Yes". (A) All the measurements together. (B) Measurements split by prepartum and postpartum. For non-binary scales, adjusted p-values from Mann-Whitney tests are shown: these assess whether prepartum and postpartum distributions for a given non-binary scale are significantly different. (C) Measurements split by visit.

prevalence of observations either not showing depression, or showing a mild level of it. Only 10% of all observations yielded depression according to EPDS (score >9), 5% according to HDRS (score >7), 20% according to MADRS (score >6), and only 6% confirmed a depressive episode according to MINI.

We assessed differences between prepartum and postpartum distributions of non-binary scales by means of Mann-Whitney tests. In two cases, namely for EPDS and HDRS, the tests revealed significant differences, with adjusted p-values of ~10^(-11) and ~10^(-3), respectively. Such differences also emerge from the boxplots of Fig. 2B, where distributions for the two periods are shown. For both EPDS and HDRS postpartum observations appear to have (in general) lower values than prepartum ones, while this difference is less substantial using MADRS and VAS tools.

3.2. *Agreement among scales during pregnancy and the postpartum*

We considered all possible pairs of non-binary scales and we studied their joint distributions, by separating prepartum and postpartum periods. Qualitative observations on differences between prepartum and postpartum and on the agreement among scales can be obtained from the scatterplots displayed in Fig. 3A. Here, by agreement we mean monotonicity to each other: two scales are in good agreement if their scatter plot shows a monotonic upward trend. In order to quantitatively support such observations, we calculated Spearman's rank correlation coefficients between all possible pairs of scales, by keeping separation between prepartum and postpartum observations. These measures of agreement are presented in the diagrams represented in Fig. 3B. The main considerations that can be derived are the following:

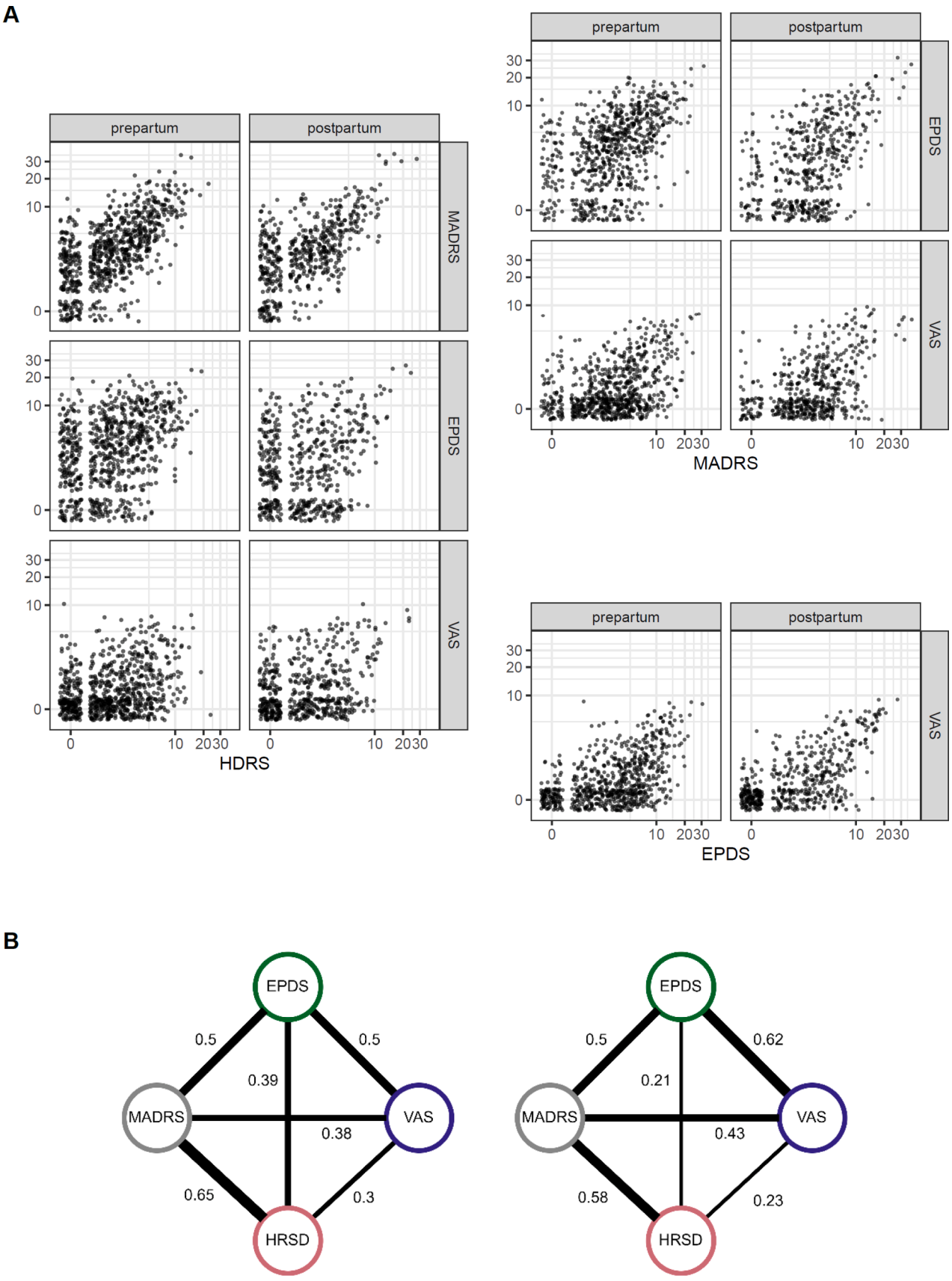

Fig. 3. (A) Scatterplots in log-log scale for pairs of non-binary scales. A jitter (0.25 on log-transformed values) was added on both axes in order to make the graph more readable. Prepartum and postpartum visits are shown in different panels. (B) Diagrams showing Spearman's correlation coefficients for pairs of non-binary scales in prepartum and postpartum.

- HDRS and MADRS show good agreement, especially for values larger than zero. In fact, the correlation coefficients are among the largest (0.65 in prepartum and 0.58 in postpartum)
- HDRS and VAS show poor agreement, especially for low values. In fact, the correlation coefficients are among the smallest (0.3 and 0.23).
- HDRS and EPDS show poor agreement, especially in postpartum. This is reflected in a decrease of the correlation coefficient from 0.39 to 0.21.
- EPDS and VAS show good agreement in general. In particular, there is more agreement on low values between the two in postpartum than during pregnancy. This is reflected in an increase of the correlation coefficient from 0.5 to 0.62.

3.3. Prediction of a binary scale using others

As a preliminary study, we performed Mann-Whitney tests to assess whether the distribution of the non-binary scales is different between women with positive or negative MINI and between those with EPDS≤9 or EPDS>9. In all cases but one (values of HDRS on samples determined by MINI), significant differences emerged (Fig. 4A). In general, the distributions for samples determined by values of binary EPDS appear more separated than for those determined by MINI. This can be appreciated in the boxplots depicted in Fig. 4B, where boxes are well apart for binary EPDS.

We trained multivariable classification models to predict both MINI and binary EPDS, using all the other available scales as features. As shown in Table 1, we obtained a well-performing model for the prediction of binary EPDS, and a less performing one for the prediction of MINI. Then, both for the prediction of MINI and binary EPDS, we trained a model for each possible subset of the available scales. In both cases, we identified a model which only considers two scales but performs comparably to the one using all the available scales. In particular, for predicting binary EPDS it is the model using only VAS and MADRS, for predicting MINI the one using only HDRS and VAS.

When dealing with imbalanced datasets, the AUC score can give overly optimistic results. In our case, this is particularly true for the MINI model, which had only 6% depressed women compared to the EPDS model's 10%. As a result, the distortion caused by the imbalance is likely to have a greater impact on the MINI model than on the EPDS model. This reinforces our conclusion that the EPDS model is superior to the

Table 1
Performances of classification models for prediction of binary EPDS and MINI. The first two rows refer to models using all available scales, the second two to models using only two scales. All models were trained in five-fold cross validation for ten repetitions. Average and standard deviation across repetitions for the following performance metrics are shown: area under receiving operating curve (AUROC), area under the precision recall curve (AUPR), sensitivity (SEN) and specificity (SPE).

Target	Scales used	AUROC	AUPR	SEN	SPE
Binary EPDS	HDRS, MADRS, VAS, MINI	0.899 ± 0.001	0.631 ± 0.003	0.801 ± 0.015	0.857 ± 0.013
MINI	HDRS, MADRS, EPDS, VAS	0.660 ± 0.009	0.174 ± 0.016	0.580 ± 0.027	0.677 ± 0.012
Binary EPDS	MADRS, VAS	0.901 ± 0.001	0.633 ± 0.005	0.810 ± 0.012	0.866 ± 0.007
MINI	HDRS, VAS	0.648 ± 0.004	0.173 ± 0.013	0.565 ± 0.022	0.637 ± 0.010

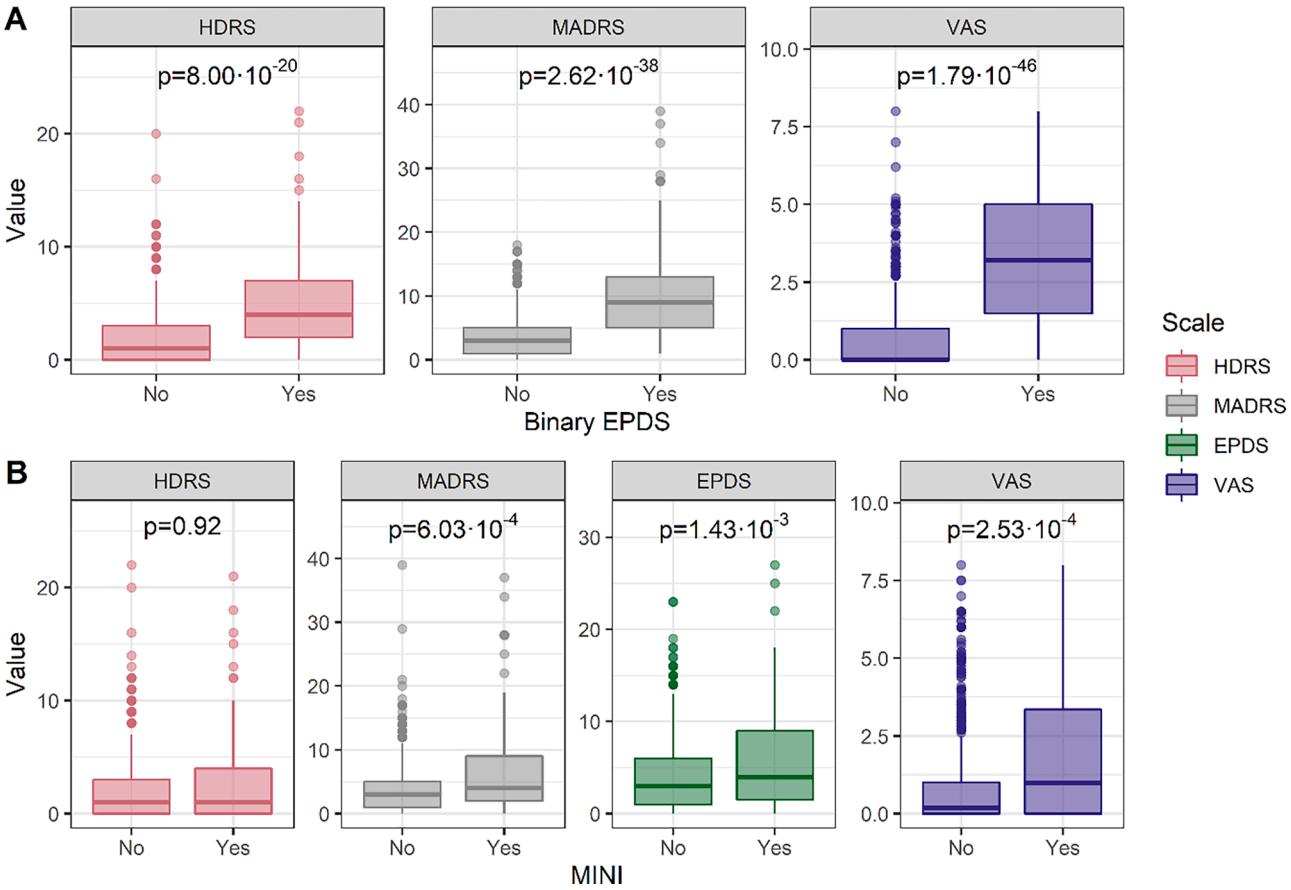

Fig. 4. (A) Boxplots showing distributions of non-binary scales for the two samples determined by dividing observations into two groups according to their value of binary EPDS. Adjusted p-values from Mann-Whitney tests are shown: these assess whether the two samples are significantly different. (B) Analogous to subfigure A, but for samples determined by means of MINI.

MINI model.

Fig. 5 illustrates repetition one (out of ten) of the classification models in major detail. Fig. 5A shows ROC and PR curves and their underlying area both for the models using all the available scales and the aforementioned models using only two scales. Fig. 5B provides confusion matrices for the models using all scales: these were determined by using the optimal threshold in terms of geometric mean of sensitivity and specificity.

4. Discussion

4.1. Multiscale prediction of PND risk (EPDS) and diagnosis (MINI)

The main finding of our study is that machine learning models employing several self- and clinician-administered depression scales classify women at risk for PND (EPDS total score > 9) substantially better than women with a MINI-confirmed Major Depressive Episode. Indeed, symptom questionnaires are not designed to ascertain diagnostic status, and our models confirmed their relatively low reliability. This finding has implications for both research and clinical practice. Symptom screening tools are often employed in both settings because administration of diagnostic interviews is time- and resource-consuming

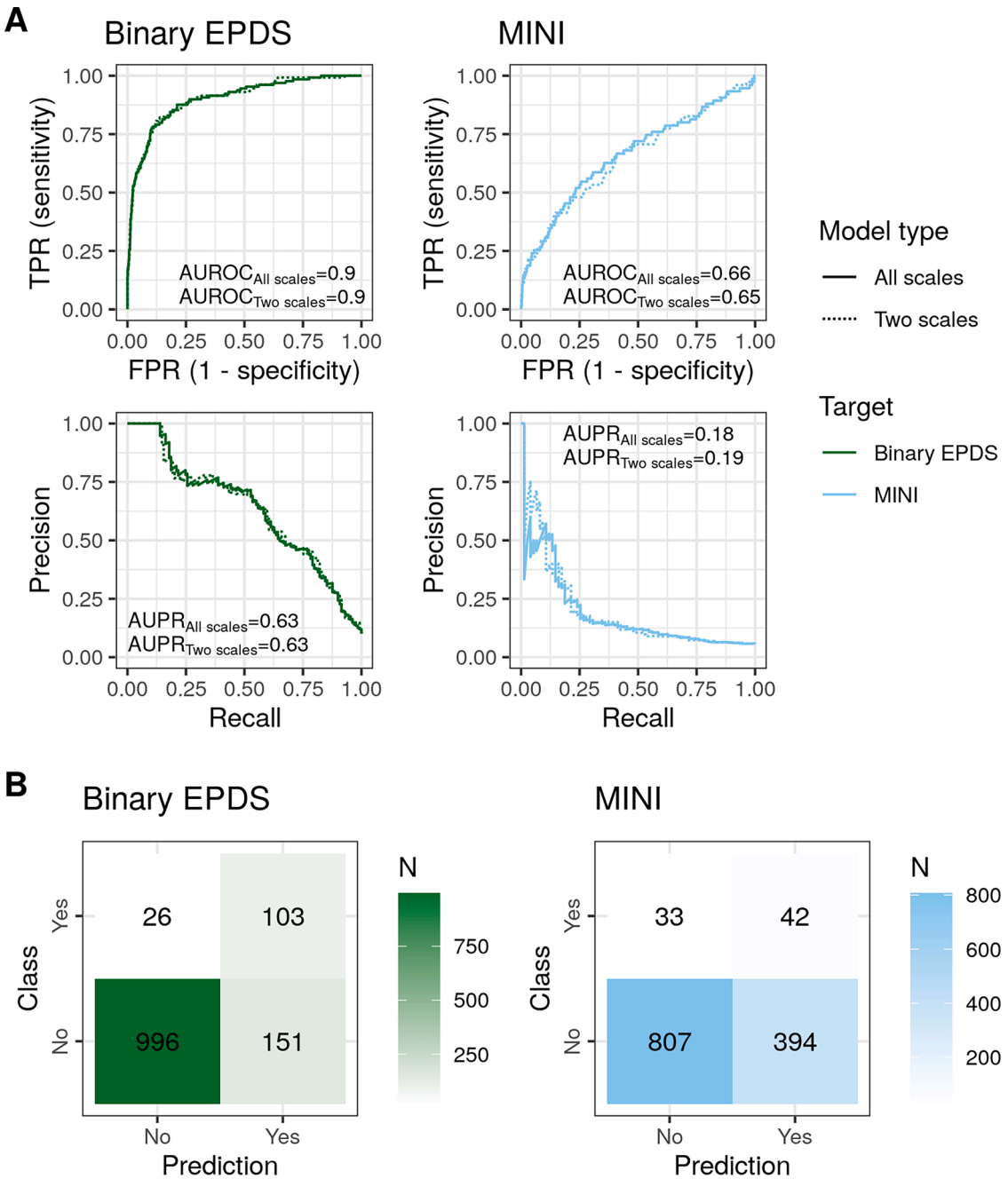

Fig. 5. (A) ROC and PR curves, with their underlying areas, for repetition one out of ten of models predicting binary EPDS (left) and MINI (right). In both cases a model using all available scales and one using only two scales are shown: for binary EPDS the two-scale one uses VAS and MADRS, for MINI it uses HDRS and VAS. (B) Confusion matrices for repetition one of models using all available scales. The threshold used for assigning a class to prediction is the one maximizing the geometric mean of sensitivity and specificity.

(Levis et al., 2018). However, all self-report and clinician-administered questionnaires appear to clearly identify PND likelihood rather than MDE diagnosis. According to some authors screening for PND through available instruments fails to confer benefits above usual clinical care, i.e. the clinician's inquiry and attention to mental health and well-being during pregnancy and postpartum (Lang et al., 2022). Given the known specificity of EPDS for the perinatal period, our findings could also suggest that the PND construct does not fully overlap with MDE. MINI might fail to capture core aspects of PND, given that previous research has shown that the likelihood of being diagnosed a MDE increases less for the MINI than for the Structured Clinical Interview for DSM Disorders (SCID) as EPDS total score increases (Levis et al., 2020). The ideal PND screening tool should limit the weight of symptoms that overlap with physiological postpartum experiences, and capture unique symptoms that are missing from MDE screening tools (Batt et al., 2020). Indeed, available instruments fail to clearly differentiate PND from the common experience of "baby blues", and from other clinical conditions with overlapping symptoms, such as generalized anxiety disorder, obsessive-compulsive disorder, and postpartum psychosis (Kettunen et al., 2014).

4.2. Agreement across psychometric instruments during pregnancy and postpartum

HDRS total score was relatively consistent with MADRS, but not with VAS nor EPDS, especially postpartum. EPDS appeared to be relatively more consistent with VAS, especially when scores were low and postpartum. The low degree of correspondence between HDRS and VAS is not surprising given their very different structure: the former requires an objective assessment of cognitive, affective and neurovegetative symptoms of depression, whereas the latter only measures the subjective experience of "feeling depressed". Indeed, the MADRS appears to correlate more closely with VAS, perhaps due to its focus on "core" depressive symptoms.

HDRS is widely known for its focus on somatic and neurovegetative symptoms of depression (Gibbons et al., 1993; Nixon et al., 2020; Vindbjerg et al., 2019). In the context of PND, HDRS may detect somatic symptoms which are common during pregnancy and/or the postpartum period (i.e. fatigue, body aches, reduced libido, sleep disturbances) but may not necessarily stem from a depressive condition and may not be associated with a subjective experience of negative affect. Hence, HDRS may be less specific and its scores may be inflated when compared to other measures. On the other hand, it may be argued that some women who develop PND may not recognize such symptoms as indicators of depression, but rather attribute them to the physiological stress of the perinatal period. The symptomatic overlap between depression and common physical complaints in pregnancy and after delivery has been previously highlighted by Ross et al., 2003. In their cohort of 150 women followed-up between 36 weeks gestation and 16 weeks postpartum, somatic item scores did not correlate with total HDRS score during pregnancy, but increased at 6 weeks postpartum, when mood items score correlation with total score lowered in comparison to pregnancy. The authors concluded that women may be more inclined to identify their complaints as physical rather than mood-related after childbirth, compared to pregnancy (Ross et al., 2003). The abundance of somatic items on HDRS might therefore act as a confounding factor in the screening process and in the assessment of severity in PND. Indeed, agreement between HDRS and EPDS strongly decreased in our cohort in postpartum observations.

We found major agreement on low values between EPDS and VAS in postpartum rather than during pregnancy, suggesting VAS may be employed as a fast screening tool after childbirth. Visual Analog Scales are straightforward, graphical self-reports of emotional states that can overcome linguistic barriers and have been employed in studies of both postpartum blues and depression (Cox et al., 1983; Kendell et al., 1981). Originally developed to assess mood in patients with neurological disturbances such as aphasia or stroke (Stern, 1997), VAS has also been employed in other clinical settings as a screening tool (Bennett et al., 2006). However, the broad range of concurrent validity coefficients (0.12–0.82) limits the interpretation of results and has raised concerns over the scales' psychometric quality in terms of validity and reliability (Athanasou, 2019).

4.3. Study limitations

Some limitations of our work must be considered. First of all, our findings might be influenced by the cut-off choice of 9 for EPDS, as higher scores have been shown to progressively yield more cases of MINI-confirmed MDE diagnoses (Levis et al., 2020). However, our choice was driven by the observation that lower cut-offs are most efficiently employed to avoid false negatives and identify most patients who meet diagnostic criteria (Levis et al., 2019a). In addition, from a purely numerical perspective, employing a cut-off value of 12 for EPDS would have resulted in a dataset with a degree of imbalance around 3.6%. Given the limited amount of data, we believe this level of imbalance would be too severe to produce reliable results. Second, relatively low rates of depression risk and MDE were found in our cohort (10% and 6% of all observations, respectively), thus limiting the overall number of positive cases in the binary EPDS and MINI prediction models. This is likely to depend on our choice to exclude women diagnosed with a depressive episode or bipolar disorder at baseline, which has been explained elsewhere (Baiardi et al., 2016). Finally, limited and varying evidence of validity in the identification of antepartum depression has been reported for EPDS (Owora et al., 2016), although it remains the most commonly used instrument in clinical practice.

3. Conclusion

Globally, our findings suggest that results derived from different scales should be compared with great caution, due to a substantial variability across women with low/high symptom scores and during pregnancy or the postpartum period. Whenever the EPDS cannot be employed, the VAS can be reliably administered for ultrarapid, extensive postpartum screening. On the other hand, commonly employed clinician-administered or self-report tools cannot reliably replace a full structured interview or the clinical examination required to establish a diagnosis of MDE.

Financial support

The"Life-ON" study was funded by the Swiss National Science Foundation (grant: 320030_160250/1) and the Italian Ministry of Health and Emilia-Romagna Region (grant: PE-2011-02348727).

CRediT authorship contribution statement

Armando D'Agostino: Conceptualization, Writing – original draft, Writing – review & editing. **Corrado Garbazza:** Conceptualization, Writing – review & editing. **Daniele Malpetti:** Data curation, Formal analysis, Methodology. **Laura Azzimonti:** Methodology, Supervision. **Francesca Mangili:** Methodology, Supervision. **Hans-Christian Stein:** Data curation, Writing – review & editing. **Renata del Giudice:** Data curation, Methodology, Writing – review & editing. **Alessandro Cicolin:** Supervision, Writing – review & editing. **Fabio Cirignotta:** Funding acquisition, Supervision, Writing – review & editing. **Mauro Manconi:** Conceptualization, Project administration, Supervision.

Declaration of Competing Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Acknowledgments

The authors thank all women and personnel who participated in the study across sites.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.psychres.2023.115687.

References

American College of Obstetricians and Gynecologists, 2018. Screening for perinatal depression. ACOG committee opinion no. 757. Obstet. Gynecol. 132, e208–e212.

Athanasou, J.A., 2019. The background, psychometric qualities and clinical application of the visual analog mood scales: a review and evaluation. Psychol. Thought 12, 265–276.

Baiardi, S., Cirignotta, F., Cicolin, A., Garbazza, C., D'Agostino, A., Gambini, O., Giordano, A., Canevini, M., Zambrelli, E., Marconi, A.M., Mondini, S., Borgwardt, S., Cajochen, C., Rizzo, N., Manconi, M, 2016. Chronobiology, sleep-related risk factors and light therapy in perinatal depression: the "Life-ON" project. BMC Psychiatry 16 (1), 374. https://doi.org/10.1186/s12888-016-1086-0. Nov 4PMID: 27814712; PMCID: PMC5225570.

Batt, M.M., Duffy, K.A., Novick, A.M., Metcalf, C.A., Epperson, C.N., 2020. Is postpartum depression different from depression occurring outside of the perinatal period? A review of the evidence. Focus 18 (2), 106–119. https://doi.org/10.1176/appi.focus.20190045 (Am Psychiatr Publ)AprEpub 2020 Apr 23. PMID: 33162848; PMCID: PMC7587887.

Benjamini, Y., Hochberg, Y., 2019. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x. https://www.jstor.org/stable/2346101.

Bennett, H.E., Thomas, S.A., Austen, R., Morris, A.M., Lincoln, N.B, 2006. Validation of screening measures for assessing mood in stroke patients. Br. J. Clin. Psychol. 45 (Pt 3), 367–376. https://doi.org/10.1348/014466505x58277. SepErratum in: Br J Clin Psychol. 2007 Jun;46(Pt 2):following 251. PMID: 17147102.

Benvenuti, P., Ferrara, M., Niccolai, C., Valoriani, V., Cox, J.L., 1999. The Edinburgh postnatal depression scale: validation for an Italian sample. J. Affect. Disord. 53 (2), 137–141. https://doi.org/10.1016/s0165-0327(98)00102-5. MayPMID: 10360408.

Bhat, A., Nanda, A., Murphy, L., Ball, A.L., Fortney, J., Katon, J., 2022. A systematic review of screening for perinatal depression and anxiety in community-based settings. Arch. Womens Ment. Health 25 (1), 33–49. https://doi.org/10.1007/s00737-021-01151-2. FebEpub 2021 Jul 11. PMID: 34247269.

Carpiniello, B., Pariante, C.M., Serri, F., Costa, G., Carta, M.G., 1997. Validation of the edinburgh postnatal depression scale in Italy. J. Psychosom. Obstet. Gynaecol. 18 (4), 280–285. https://doi.org/10.3109/01674829709080700. DecPMID: 9443138.

Cellini, P., Pigoni, A., Delvecchio, G., Moltrasio, C., Brambilla, P., 2022. Machine learning in the prediction of postpartum depression: a review. J. Affect. Disord. 309, 350–357. https://doi.org/10.1016/j.jad.2022.04.093. Jul 15Epub 2022 Apr 20. PMID: 35460742.

Cox, J., 2019. Thirty years with the Edinburgh postnatal depression scale: voices from the past and recommendations for the future. Br. J. Psychiatry 214 (3), 127–129. https://doi.org/10.1192/bjp.2018.245. MarPMID: 30774059.

Cox, J.L., Connor, Y.M., Henderson, I., McGuire, R.J., Kendell, R.E., 1983. Prospective study of the psychiatric disorders of childbirth by self report questionnaire. J. Affect. Disord. 5 (1), 1–7. https://doi.org/10.1016/0165-0327(83)90030-7. FebPMID: 6220039.

Cuijpers, P., Franco, P., Ciharova, M., Miguel, C., Segre, L., Quero, S., Karyotaki, E., 2021. Psychological treatment of perinatal depression: a meta-analysis. Psychol. Med. 1–13.

Gibbons, R.D., Clark, D.C., Kupfer, D.J., 1993. Exactly what does the hamilton depression rating scale measure? J. Psychiatr. Res. 27 (3), 259–273. https://doi.org/10.1016/0022-3956(93)90037-3.

First, M.B., Williams, J.B.W., Karg, R.S., Spitzer, R.L., 2016. Structured Clinical Interview for DSM-5 Disorders, Clinician Version (SCID-5-CV). American Psychiatric Association, Arlington, VA.

Garbazza, C., Cirignotta, F., D'Agostino, A., Cicolin, A., Hackethal, S., Wirz-Justice, A., Cajochen, C., M;, Manconi, 2022. Life-ON" study group. Sustained remission from perinatal depression after bright light therapy: a pilot randomised, placebo-controlled trial. Acta Psychiatr. Scand. https://doi.org/10.1111/acps.13482. Jul 25Epub ahead of print. PMID: 35876837.

Garbazza C., Mangili F., D'Onofrio T.A., Malpetti D., Riccardi S., Cicolin A., D'Agostino A., Cirignotta F., Manconi M., and the "Life-ON" study group. A machine learning model for predicting the risk of perinatal depression in pregnant women. Under review.

Hamilton, M., 1960. A rating scale for depression. J. Neurosurg. Psychiatry 23, 56–62. https://doi.org/10.1136/jnnp.23.1.56.

Kendell, R.E., McGuire, R.J., Connor, Y., Cox, J.L., 1981. Mood changes in the first three weeks after childbirth. J. Affect. Disord. 3 (4), 317–326. https://doi.org/10.1016/0165-0327(81)90001-x. DecPMID: 6459348.

Kettunen, P., Koistinen, E., Hintikka, J., 2014. Is postpartum depression a homogenous disorder: time of onset, severity, symptoms and hopelessness in relation to the course of depression. BMC Pregnancy Childbirth 14, 402. https://doi.org/10.1186/s12884-014-0402-2.

Kuhn M. Caret: classification and regression training. R package version 6.0-90. https://cran.r-project.org/web/packages/caret/caret.pdf; 2021 [Accessed 16 August 2022].

Lang, E., Colquhoun, H., LeBlanc, J.C., Riva, J.J., Moore, A., Traversy, G., Wilson, B., Grad, R., 2022. Canadian task force on preventive health care. Recommendation on instrument-based screening for depression during pregnancy and the postpartum period. CMAJ 194 (28), E981–E989. https://doi.org/10.1503/cmaj.220290. Jul 25PMID: 35878894; PMCID: PMC9328462.

Lecrubier, Y., Sheehan, D.V., Weiller, E., Amorim, P., Bonora, I., Sheehan, K.H., et al., 1997. The mini international neuropsychiatric interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI. Eur. Psychiatry 12, 224–231.

Levis, B., Benedetti, A., Riehm, K.E., Saadat, N., Levis, A.W., Azar, M., Rice, D.B., Chiovitti, M.J., Sanchez, T.A., Cuijpers, P., Gilbody, S., Ioannidis, J.P.A., Kloda, L.A., McMillan, D., Patten, S.B., Shrier, I., Steele, R.J., Ziegelstein, R.C., Akena, D.H., Arroll, B., Ayalon, L., Baradaran, H.R., Baron, M., Beraldi, A., Bombardier, C.H., Butterworth, P., Carter, G., Chagas, M.H., Chan, J.C.N., Cholera, R., Chowdhary, N., Clover, K., Conwell, Y., de Man-van Ginkel, J.M., Delgadillo, J., Fann, J.R., Fischer, F.H., Fischler, B., Fung, D., Gelaye, B., Goodyear-Smith, F., Greeno, C.G., Hall, B.J., Hambridge, J., Harrison, P.A., Hegerl, U., Hides, L., Hobfoll, S.E., Hudson, M., Hyphantis, T., Inagaki, M., Ismail, K., Jetté, N., Khamseh, M.E., Kiely, K.M., Lamers, F., Liu, S.I., Lotrakul, M., Loureiro, S.R., Löwe, B., Marsh, L., McGuire, A., Mohd Sidik, S., Munhoz, T.N., Muramatsu, K., Osório, F.L., Patel, V., Pence, B.W., Persoons, P., Picardi, A., Rooney, A.G., Santos, I.S., Shaaban, J., Sidebottom, A., Simning, A., Stafford, L., Sung, S., Tan, P.L.L., Turner, A., van der Feltz-Cornelis, C.M., van Weert, H.C., Vöhringer, P.A., White, J., Whooley, M.A., Winkley, K., Yamada, M., Zhang, Y., Thombs, B.D, 2018. Probability of major depression diagnostic classification using semi-structured versus fully structured diagnostic interviews. Br. J. Psychiatry 212 (6), 377–385. https://doi.org/10.1192/bjp.2018.54. JunEpub 2018 May 2. PMID: 29717691; PMCID: PMC6415695.

Levis, B., McMillan, D., Sun, Y., He, C., Rice, D.B., Krishnan, A., Wu, Y., Azar, M., Sanchez, T.A., Chiovitti, M.J., Bhandari, P.M., Neupane, D., Saadat, N., Riehm, K.E., Imran, M., Boruff, J.T., Cuijpers, P., Gilbody, S., Ioannidis, J.P.A., Kloda, L.A., Patten, S.B., Shrier, I., Ziegelstein, R.C., Comeau, L., Mitchell, N.D., Tonelli, M., Vigod, S.N., Aceti, F., Alvarado, R., Alvarado-Esquivel, C., Bakare, M.O., Barnes, J., Beck, C.T., Bindt, C., Boyce, P.M., Bunevicius, A., Couto, T.C.E., Chaudron, L.H., Correa, H., de Figueiredo, F.P., Eapen, V., Fernandes, M., Figueiredo, B., Fisher, J.R.W., Garcia-Esteve, L., Giardinelli, L., Helle, N., Howard, L.M., Khalifa, D.S., Kohlhoff, J., Kusminskas, L., Kozinszky, Z., Lelli, L., Leonardou, A.A., Lewis, B.A., Maes, M., Meuti, V., Nakić Radoš, S., Navarro García, P., Nishi, D., Okitundu Luwa, E., Andjafono, D., Robertson-Blackmore, E., Rochat, T.J., Rowe, H.J., Siu, B.W.M., Skalkidou, A., Stein, A., Stewart, R.C., Su, K.P., Sundström-Poromaa, I., Tadinac, M., Tandon, S.D., Tendais, I., Thiagayson, P., Töreki, A., Torres-Giménez, A., Tran, T.D., Trevillion, K., Turner, K., Vega-Dienstmaier, J.M., Wynter, K., Yonkers, K.A., Benedetti, A., Thombs, B.D, 2019a. Comparison of major depression diagnostic classification probability using the SCID, CIDI, and MINI diagnostic interviews among women in pregnancy or postpartum: an individual participant data meta-analysis. Int. J. Methods Psychiatr. Res. 28 (4), e1803. https://doi.org/10.1002/mpr.1803. DecEpub 2019 Sep 30. PMID: 31568624; PMCID: PMC7027670.

Levis, B., Negeri, Z., Sun, Y., Benedetti, A., Thombs, B.D., 2020. Depression screening data (DEPRESSD) EPDS group. 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, m4022. https://doi.org/10.1136/bmj.m4022. Nov 11PMID: 33177069; PMCID: PMC7656313.

Levis, B., Yan, X.W., He, C., Sun, Y., Benedetti, A., Thombs, B.D., 2019b. Comparison of depression prevalence estimates in meta-analyses based on screening tools and rating scales versus diagnostic interviews: a meta-research review. BMC Med. 17 (1), 65. https://doi.org/10.1186/s12916-019-1297-6. Mar 21PMID: 30894161; PMCID: PMC6427845.

Mitchell, P.B., Frankland, A., Hadzi-Pavlovic, D., Roberts, G., Corry, J., Wright, A., Loo, C.K., Breakspear, M, 2011. Comparison of depressive episodes in bipolar disorder and in major depressive disorder within bipolar disorder pedigrees. Br. J. Psychiatry 199 (4), 303–309. https://doi.org/10.1192/bjp.bp.110.088823. OctEpub 2011 Apr 20. PMID: 21508436.

Montgomery, S.A., Asberg, M., 1979. A new depression scale designed to be sensitive to change. Br. J. Psychiatry 134, 382–389. https://doi.org/10.1192/bjp.134.4.382. AprPMID: 444788.

Moussavi, S., Chatterji, S., Verdes, E., Tandon, A., Patel, V., Ustun, B., 2007. Depression, chronic diseases and decrements in health: results from the word health surveys. Lancenet 370 (9590), 851–858. https://doi.org/10.1016/S0140-6736(07)61415-9.

Nixon, N., Guo, B., Garland, A., Kaylor-Hughes, C., Nixon, E., Morriss, R., 2020. The Bi-factor structure of the 17-item Hamilton depression rating scale in persistent major depression; dimensional measurement of outcome. PLoS One 15 (10), e0241370. https://doi.org/10.1371/journal.pone.0241370. Oct 26PMID: 33104761; PMCID: PMC7588071.

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. J. Affect. Disord. 205, 335–343.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ [Accessed 16 August 2022].

Reps, J.M., Wilcox, M., McGee, B.A., Leonte, M., LaCross, L., Wildenhaus, K., 2022. Development of multivariable models to predict perinatal depression before and

after delivery using patient reported survey responses at weeks 4-10 of pregnancy. BMC Pregnancy Childbirth 22 (1), 442. https://doi.org/10.1186/s12884-022-04741-9. May 26PMID: 35619056; PMCID: PMC9137134.

Ross, L.E., Gilbert Evans, S.E., Sellers, E.M., Romach, M.K, 2003. Measurement issues in postpartum depression part 1: anxiety as a feature of postpartum depression. Arch. Womens Ment. Health 6 (1), 51–57. https://doi.org/10.1007/s00737-002-0155-1. FebPMID: 12715264.

Sheehan, D.V., Lecrubier, Y., Sheehan, K.H., Janavs, J., Weiller, E., Keskiner, A., Schinka, J., Knapp, E., Sheehan, M.F., Dunbar, G.C., 1997. The validity of the mini international neuropsychiatric interview (MINI) according to the SCID-P and its reliability. Eur. Psychiatry 12 (5), 232–241. https://doi.org/10.1016/S0924-9338(97)83297-X.

Snaith, R.P., 1996. Present use of the Hamilton depression rating scale: observation on method of assessment in research of depressive disorders. Br. J. Psychiatry 168 (5), 594–597. https://doi.org/10.1192/bjp.168.5.594. MayPMID: 8733798.

Sockol, L.E., Epperson, C.N., Barber, J.P., 2011. A meta-analysis of treatments for perinatal depression. Clin. Psychol. Rev. 31 (5), 839–849. https://doi.org/10.1016/j.cpr.2011.03.009. JulEpub 2011 Mar 27. PMID: 21545782; PMCID: PMC4108991.

Stern, R.A., 1997. Visual Analog Mood Scales. Professional Manual. Psychological Assessment Resources, Lutz, FL, USA.

Vindbjerg, E., Makransky, G., Mortensen, E.L., Carlsson, J., 2019. Cross-cultural psychometric properties of the hamilton depression rating scale. Can. J. Psychiatry 64 (1), 39–46. https://doi.org/10.1177/0706743718772516. JanEpub 2018 May 2. PMID: 29719964; PMCID: PMC6364134.

Wakefield, C., Frasch, M.G., 2023. Predicting patients requiring treatment for depression in the postpartum period using common electronic medical record data available antepartum. AJPM Focus 2 (3), 100100. https://doi.org/10.1016/j.focus.2023.100100. Apr 27PMID: 37790672; PMCID: PMC10546501.

Woody, C.A., Ferrari, A.J., Siskind, D.J., Whiteford, H.A., Harris, M.G., 2017. A systematic review and meta-regression of the prevalence and incidence of perinatal depression. J. Affect. Disord. 219, 86–92. https://doi.org/10.1016/j.jad.2017.05.003. SepEpub 2017 May 8. PMID: 28531848.

Xu, W., Sampson, M., 2023. Prenatal and childbirth risk factors of postpartum pain and depression: a machine learning approach. Matern. Child Health J. 27 (2), 286–296. https://doi.org/10.1007/s10995-022-03532-0. FebEpub 2022 Dec 16. PMID: 36526882.

Yu, Z., Matsukawa, N., Saigusa, D., Motoike, I.N., Ono, C., Okamura, Y., Onuma, T., Takahashi, Y., Sakai, M., Kudo, H., Obara, T., Murakami, K., Shirota, M., Kikuchi, S., Kobayashi, N., Kikuchi, Y., Sugawara, J., Minegishi, N., Ogishima, S., Kinoshita, K., Yamamoto, M., Yaegashi, N., Kuriyama, S., Koshiba, S., Tomita, H., 2022. Plasma metabolic disturbances during pregnancy and postpartum in women with depression. iScience 25 (12), 105666. https://doi.org/10.1016/j.isci.2022.105666. Nov 24PMID: 36505921; PMCID: PMC9732390.

Zhong, M., Zhang, H., Yu, C., Jiang, J., Duan, X., 2022. Application of machine learning in predicting the risk of postpartum depression: a systematic review. J. Affect. Disord. 318, 364–379. https://doi.org/10.1016/j.jad.2022.08.070. Dec 1Epub 2022 Aug 31. PMID: 36055532.