*Martin Bernstorff | #201307263Supervisor: Anders Foldspang*

*18/10-16*

A comparison of the Edinburgh Postnatal Depression Scale (EPDS) and the Postpartum Depression Screening Scale (PDSS) for peripartum depression screening

# Bachelorprojekt

# Abstract

# Problem Statement

· Does EPDS or PDSS offer the largest area under the curve in a receiver-operating-characteristics-curve?

· What are the trade-offs in deciding on an appropriate cut-off value for each questionnaire in this setting?

# 1. Introduction

*The basic background to the question you will work with, ending with a  brief and clear statement of the aim of your work, one aim being better than more aims (!). In this section you may cite individual articles, reviews and other (hopefully) reliable sources (e.g. textbooks). Brevity and clarity are basic virtues.*

Major depressive disorder (MDD)

Major peri-partum depression (MPPD)

Minor peri-partum depression (mPPD)

Postpartum Depression Screening Scale (PDSS)

Edinburg Postpartum Depression Scale (EPDS)

Diagnostic and Statistical Manual of Mental Disorders (DSM)

Receiver-operator characteristics (ROC)

# 2. Methods

**Search string**

("screening”[title] AND (“EPDS” OR “Edinburgh Postnatal Depression Scale”) AND ("Postpartum Depression Screening Scale” OR “PDSS”))

AND

(“comparative study”[publication type] OR “combined”[title] OR “comparison”[title] OR “comparative”[title])

AND

(“sensitivity” OR “specificity”)

NOT

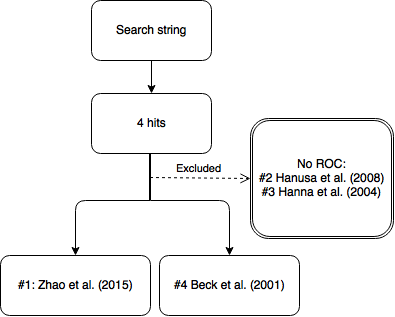
(“review”[publication type])

This is a Bachelor’s thesis and must therefore be written within certain boundaries. For this reason, the search-scope has been narrowed extensively.

“Pga. vejleder […], ellers ville jeg”

Searches were performed in the PubMed database.

The search-string consists of 4 blocks.

1. Subject matter.
2. Only comparative studies, as to isolate the characteristics of the questionnaires. Comparing the questionnaires via studies with information on only one questionnaire would run the risk of comparing the demographics of the studies, not the qualities of the questionnaires.
3. Studies must use the word sensitivity or specificity in their abstract, to increase the chance of them supplying it in the article.
4. Reviews were excluded as we were instructed to use only original articles.

This search string returns 4 hits. Articles that did not contain ROC-curves (n = 2) were excluded.

This leaves us with two suitable articles, Zhao et al. and Beck et al.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Criterion** | **Beck et al. (2001)** | | **Zhao et al. (2015)** | |
| **Sample size** | 150 | | 842 | |
| **Inclusion criteria** | · Age ≥ 18  · Able to speak and read English  · 2-12 weeks post-partum  · Delivered a live, healthy infant | | · Obstetric complication  · Pregnant | |
| **Reference-standard test** | DSM-IV diagnostic interview | | Not applicable | |
| **Country** | United States | | China | |
| **Language** | English | | Chinese | |
|  | **EPDS** | **PDSS** | **EPDS** | **PDSS** |
| **Cut-off (MPPD)** | 12/13 | 79/80 | 12/13 | 79/80 |
| **Cronbach’s α(entire test)** | 0.89 | Not reported | 0.78 | 0.95 |
| **AUC (MPPD)** | 0.96 | 0.98 | 0.898\* | 0.983\* |
| **AUC (MPPD & mPPD)** | 0.83\* | 0.91\* | 0.822\* | 0.979\* |
| **Interviewer** | Nurse psychotherapist | | Trained research assistant | |
| **Blinding** | Yes (interviewer blind to scores) | | No (only high-risk women interviewed) | |

# 3. Findings

*\* EPDS vs. PDSS statistically significant (p < 0.001)*

*Focus areas from the aim statement are investigated in depth based upon the findings of original research articles. Keep the strict connection to the aim(s)!*

Blablabl

# 4. Discussion

For a comparison of articles to make sense, the methodology of the articles must be adequately similar.

**4.1 Test protocol**

**4.1.1 Construct**

*4.1.1.1 Construct similarity*

Beck et al. examine postpartum whereas Zhao et al. examine antepartum. In the DSM-V, depressive disorders can be appended the qualifier ‘with peripartum onset’ if manifestation is during pregnancy or in the 4 weeks following birth1. Following this example, ante- and postpartum depression are not examined as two separate constructs in this thesis.

The PDSS is made specifically for post-partum depression, as seen in some of its questions: “I had trouble sleeping even when my baby was asleep.”2 This question makes no sense in the context of antepartum depression, Presumably, Zhao et al. must have modified this question. No such information is given in Zhao et al.

Beck et al. published their article before the publishing of the DSM-V. A natural concern is that their diagnostic criteria for depression would be different than the ones of Zhao et al. However, both articles use a semi-structured interview with the diagnostic criteria of the DSM-IV as their reference-standard. Secondly, the DSM-V has seen no modifications in the criteria for depression relevant to this thesis, except that the specifier ‘with peripartum onset’ has been added3.

To imply unity among researchers around the criteria of peripartum depression would be excessive, but the working construct of this thesis and the included articles is the construct of the DSM-V.

**4.1.2 Index test**

*4.1.2.1 Index test comparability*

The comparability of the English and Chinese version of the PDSS is ensured by proper forward-backward translatability and validation4. The same holds true for the EPDS5.

**4.1.3 Reference standard**

For an analysis of a screening tool to be meaningful, a suitable reference-standard test must be used.

In the case of depression, the accepted reference-standard in the literature is a DSM-structured or semi-structured diagnostic interview. A discussion of whether this choice is valid is outside the scope of this thesis.

*4.1.3.1 Inter-observer variation*

The interview appears to have sufficient interrater reliability with Cohen’s kappas between .7 and 1 for each dimension6,7. An assessment of inter-observer variation for the present observers would have strengthened the results of the studies.

In Beck et al. the interviewer is a nurse psychotherapist. There is no explicit information on whether multiple therapists are used.

*4.1.3.1 Diagnostic review bias*

However, every interviewer can affect the interview differently. It is therefore vital that the interviewer is blind to the scores of the screening test and, in the case of multiple interviewers, their concordance is assessed. If the interviewer preferentially diagnoses patients as depressed if they scored highly on one test, this test’s sensitivity and specificity will be artificially inflated.

In beck et al., the interviewer is blind to the screening results.

*4.1.3.1 Diagnostic test comparability*

The comparability of the Chinese and English versions of the DSM-IV interview is not sufficiently accounted for. The major study validating the translation contain very few cases of depression8. This weakens a comparison of the studies.

To estimate sensitivity and specificity, information on both true negatives, true positives, false negatives and false positives must be obtained. To know false positives and negatives, all screening results must be confirmed by a reference-standard test. In Zhao et al. this is not the case, as the reference-standard test has not been administered to all participants:

“Relative to other studies that reported the sensitivities and specificities of the screening measures for postpartum depression in comparison with diagnostic instruments […] the present study combined two depression screening tools […] to determine the efficacy without comparison with any diagnostic (e.g., SCID or DIS) instruments.” (p. 117 bottom left)

While many such methods exist, they all come with different methodological considerations and should be employed when a reference-standard test is not available9. Due to Zhao et al. not reporting which method they have used, any analysis of their statistical methodology is reduced to guess-work, and one must therefore hold the conclusions of the study in very low regard.

**4.1.4 Study populations**

In general, the method of recruitment and composition of the participants will only affect the generalisability of the results, not the internal validity.

*4.1.4.1 Demographic features*

Variation in the cultural manifestations of depression can affect the difference between the results of the two studies if one test’s questions more adequately encompass the dimensions of one culture. A generalized recommendation across cultures can therefore be dubious. Given that both studies favour the PDSS, this effect does not appear to influence the comparison critically.

In Beck et al., the mothers’ mean number of days since delivery was 39 (SD = 10.67). If the distribution is approximately symmetrical around the mean, a large amount of the participants will exceed the 4-week postpartum criterion for the DSM-V specifier of peripartum depression. The reasoning for this cut-off in the DSM-V is not expanded upon, and it’s therefore hard to gauge the severity of this discrepancy. It does, however, weaken the comparability of the study to the working-construct.

81% of the women in Beck et al. held at least a college degree. This number is abnormally high for the US, where the average number is 44% for 25-29 year olds, and 42% for 25 and over10. For this to affect the comparison between the questionnaires, one questionnaire must systematically result in a different result due to the educational level of the study participants. Given that the questionnaires read at a 3rd grade level, this seems unlikely.

*4.1.4.2 Disease prevalence*

Zhao et al. exclusively examine women with obstetric complications. Obstetric complications are a stressor, but it seems unlikely that they will change the peripartum depression construct in a way that will favour either scale, seeing as neither scale contains questions regarding obstetric complications. It might, however, increase the prevalence of depression and therefore be a source of variation for sensitivity and specificity11. The mode of this association is not known to the author of this thesis, however an association has been found empirically. This variation would not bias a comparison of the two tests.

*4.1.4.3 Population size*

Neither study has done calculations on the amount of participants required to attain sufficient statistical power. Such methods are readily available12. The articles therefore run the risk of recruiting either too few or too many participants, resulting in insufficient statistical significance or an unnecessarily high cost of information, respectfully. This has manifested itself in only one AUC-comparison reaching statistical significance in Beck et al. P < 0.05 is only a guideline, however, and every AUC in Beck et al. favours the PDSS. The P-values clearly trend inversely with the number of subjects in each group, indicating that the number of subjects, rather than an underlying lack of difference in AUC, might be able to explain the P-values.

**4.1.5 Flow and timing**

*4.1.5.1 Disease progression bias (time difference between index test and reference test)*

In Beck et al., the reference-test is administered immediately following the index-test. The risk of disease-progression bias is therefore minimal.

**4.1.6 Considerations in selection of optimal cut-off value**

*4.1.6.1 Consequences of a false-positive*

The consequence of a false-positive screen in a well-managed hospital is a diagnostic interview. However, a positive screen might affect point-of-care behaviour independently of the diagnostic interview due to inappropriate labelling13. The stress imposed on the women from this is gauged to be minimal. Given the history of downplaying the consequences of a positive screen in breast-cancer screening, further research is needed.

*4.1.6.2 Consequences of a false-negative*

A false negative might falsely reassure caregivers into believing that depression is not a possibility. This can lead to a lack of appropriate treatment and therefore worse outcomes for the patient.

Deciding on the optimal cut-off score is therefore not simply a question of finding the point on a ROC-curve closes to the top-left corner, but a consideration of the above trade-offs.

For this thesis’ comparison to be improved upon, further studies are needed with better methodology consistent with the accepted practices and with sufficiently large study populations to warrant conclusions for both major and minor depression. A broader search scope would be relevant for a comprehensive meta-analysis.

# Conclusion

*Based on the aims, methods, findings and discussion, a very brief summary of the research evaluated in the project as well as an opportunity to suggest future directions for the research area you have analysed – just a few lines with statements – no discussion (!).*

# References

1. Association, A. P. & Force, D.-5. T. *Diagnostic and statistical manual of mental disorders : DSM-5.* 186–187 (dsm.psychiatryonline.org, 2013).

2. Beck, C. T. & Gable, R. K. Comparative analysis of the performance of the Postpartum Depression Screening Scale with two other depression instruments. *Nurs Res* **50,** 242–250 (2001).

3. Highlights of Changes from DSM-IV-TR to DSM-5. 1–19 (2013).

4. Li, L., Liu, F., Zhang, H., Wang, L. & Chen, X. Chinese version of the Postpartum Depression Screening Scale: translation and validation. *Nurs Res* **60,** 231–239 (2011).

5. Wang, Y. *et al.* Psychometric evaluation of the Mainland Chinese version of the Edinburgh Postnatal Depression Scale. *Int J Nurs Stud* **46,** 813–823 (2009).

6. Maffei, C. *et al.* Interrater reliability and internal consistency of the structured clinical interview for DSM-IV axis II personality disorders (SCID-II), version 2.0. *J. Pers. Disord.* **11,** 279–284 (1997).

7. What is the relaibility of the SCID-II? *scid.org* Available at: http://www.scid4.org/psychometric/scidII\_reliability.html. (Accessed: 26 October 2016)

8. So, E. *et al.* The Chinese-bilingual SCID-I/P Project: Stage 1 — Reliability for Mood Disorders and Schizophrenia. *Hong Kong Journal of Psychiatry* 7–18 (2003).

9. Reitsma, J. B., Rutjes, A. W. S., Khan, K. S., Coomarasamy, A. & Bossuyt, P. M. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *J Clin Epidemiol* **62,** 797–806 (2009).

10. **Educational Attainment in the United States: 2014**. Available at: https://www.census.gov/hhes/socdemo/education/data/cps/2014/tables.html. (Accessed: 30 October 2016)

11. Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S.QUADAS-2 Steering Group. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *J Clin Epidemiol* **66,** 1093–1104 (2013).

12. Hess, A. S. *et al.* Methods and recommendations for evaluating and reporting a new diagnostic test. *Eur. J. Clin. Microbiol. Infect. Dis.* **31,** 2111–2116 (2012).

13. Sheehan, A. M. & McGee, H. Screening for depression in medical research: ethical challenges and recommendations. *BMC Med Ethics* **14,** 4 (2013).