*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?

# 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)

# 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

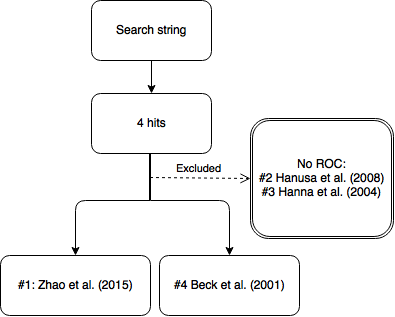
(“review”[publication type])

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 | |
| **Gold-standard test** | DSM-IV diagnostic interview | | M.I.N.I. | |
| **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) | |

# 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

# Discussion

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

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.

“One PDSS item, for example, is ‘I had trouble sleeping even when my baby was asleep.’”{Beck:2001wz} 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 gold-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 added2.

To imply unity among researchers around the criteria of peripartum depression would be excessive, but the working construct of this thesis and the articles included is the construct of the DSM-V. – Nej, Zhao et al. er før DSM-V… Beskrivelse af DSM-IV vs. V senere.

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.

For an analysis of a screening tool to be meaningful, a suitable gold-standard test must be used. In the case of depression, the gold standard is a DSM-structured or semi-structured diagnostic interview. Whether this choice is valid is outside the scope of this thesis. The interview appears to have sufficient interrater reliability with Cohen’s kappas between .7 and 1 for each dimension3,4.

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 a nurse psychotherapist. There is no explicit information on whether multiple therapists are used. The interviewer is blind to the screening results.

In Zhao et al. gold-standard testing is administered to the women who had high scores (EPDS ≥ 9 or PDSS ≥ 60). The test is the M.I.N.I. and was administered by the trained research assistant. The scale is designed for use with minimal training and the selection of interviewer is therefore appropriate. There is no explicit statement of blinding of the interviewer. Therefore, there’s a risk of misclassification. If this misclassification by the interviewer is non-random, i.e. the interviewer exhibits preference towards the results of either questionnaire, it can increase the specificity and sensitivity of this questionnaire, while decreasing the corresponding values for the other questionnaire.

fhgf

The articles employ different gold-standard tests. The MINI and DSM-IV semi-structured interviews employed by each article have been reported as comparable with a Cohen’s kappa of 0.84 for MDD5. This indicates that the gold-standard tests are comparable.

Both the language of the gold-standard test and the screening test differ between the studies. For the gold-standard tests, the Chinese version of the M.I.N.I. has been shown to correlate with the DSM-IV interview in Chinese6. However, 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 depression and a kappa of 0.5 for the inter-rater reliability of accessing moderate depression7. This weakens a comparison of the studies.

The comparability of the English and Chinese version of the PDSS is ensured by proper forward-backward translatability and validation7. The same holds true for the EPDS8.

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.

The questionnaire reliability is accessed by the mean correlation coefficient between all possible split-halves of the questionnaire, i.e. Cronbach’s α. The tests all show sufficient internal consistency in these studies without questions becoming redundant. One must be aware that the αs of two tests of differing lengths are not comparable, since increasing the amount of questions will inevitably increase Cronbach’s α9. This explains why Beck et al. have decided to publish the αs exclusively for each dimension of the PDSS, not for the entire test, – how does that explain it? Is each dimension of similar length as the PDSS?

The conclusions of the studies would have been strengthened by a test-retest to determine temporal reliability.

Neither study has done calculations on the amount of participants required to attain sufficient statistical power. They 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, may explain the P-values.

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.

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

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 gold-standard test. In Zhao et al. this is not the case, as the gold-standard test has not been administered to all participants. This seems unbelievable, but to quote the authors:

“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)

This is the extent to which their reasoning for this unorthodox method is described. This reduces any analysis of their methodology to guess-work, and therefore means that the conclusions of the study must be held in very low regard.

Part of Zhao et al.’s follow-up for women screening positive for depression is an intervention to make sure the women are “reminded to adjust the mood during pregnancy”. What this entails is not further expanded upon. This does not affect the results of the study, but the exemplification of a lack of clarity weakens our trust in the study results and description of methodology.

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

3. 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).

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

5. Sheehan, D. V. *et al.* The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry* **59 Suppl 20,** 22–33–quiz 34–57 (1998).

6. Tian-Mei, S., Dang, W.-M., Se, Y.-A. & Chen, J.-X. **Evaluation of the Reliability and Validity of Chinese Version of the Mini-International Neuropsychiatric Interview in Patients with Mental Disorders**. *Chinese Mental Health Journal* (2009).

7. 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).

8. 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).

9. Tavakol, M. & Dennick, R. Making sense of Cronbach's alpha. *Int. J. Medical Education* **2,** 53–55 (2011).

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)