PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Evaluation of the accuracy in detecting cervical lesions by nurses versus doctors using a stationary colposcope and Gynocular in a low-resource setting
AUTHORS	Wikström Shemer, Elisabet; Nessa, Ashrafun; Roy, Joya; Chowdhury, Most. Afroza; Khanam, Quayuma; Afroza, Romena; Wistrand, Charlotte; Thursesson, Marcus; Thorsell, Amlin; Shemer, Isaac

VERSION 1 - REVIEW

REVIEWER	Björn Strander, MD PhD Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden, Department of Obstetrics and Gynecology
REVIEW RETURNED	10-Jun-2014

GENERAL COMMENTS	This is an interesting study with possible great future implications. Visual inspection with acetic acid (VIA) has earlier been shown to decrease mortality in cervical cancer by 30 - 35% in a third world setting, but as a screening tool to a high extent been relying on evaluation with colposcopy, punch biopsies and laboratory facilities. Although not specifically investigating that, this study from Bangladesh, opens up the possibilities for a see an treat approach with higher specificity than VIA alone can offer, done by trained nurses, using Swede score and a portable colposcopy. A quite crucial point in this study is blinding. Obviously there is no blinding between the instruments. This was not possible, but there is a bias introduced when the examiner examines the same women twice in a row with two different instruments. The conclusion that is drawn from the first examination of course can influence the second. The block randomisation prevents this favouring one instrument over the other, but as concordance is presented the procedure will bias the results. The ideal design might be that the same examination will see the woman a second time at a randomly selected moment, being assigned the colposcope not used the first time and being blinded to her identity. This is probably practically impossible, but the limitation should be discussed. Unfortunately there is no mentioning of blinding between nurses and doctors. It appears that both were present at the same time. Were there any provisions against communicating the findings between the examiners? A study nurse "immediately recorded" the scores. Was the communication with this nurse oral or in writing? Could the other examiner present overhear it? This should be stated to assess

	the validity of the results. To assess the vessel patterns green filter without acetic acid was used. This is not in accordance with the original Swede score studies where assessment of mosaic pattern, punctuation, atypical vessels or absence of vessels was made after acetic acid application. This should be noted and discussed. No basic studies have been done so far testing Swede score on postmenopausal women or breastfeeding women with atrophy of the vaginal mucosa. The actual range of age in this study is not presented, but the data in table 1 indicate that older women participated. I recommend that results from women in postmenopausal age as well as breastfeeding women, if these can be identified, should be analysed separately. Sensitivity and specificity are stated in Results and tables and I assume this is related to find CIN2 or higher degrees of dysplasia or cancer (CIN2+) However this should be stated. The number of observations should be included in tables 2 and 3. According to the manuscript Swedescore rated by doctors ≥ 4 are included in the study, but level 4 is missing in table 3. I have great respect for the substantial problems in setting up a large study like this in this third world setting, and the number of women included is quite an achievement. This was of course limited by what was practically possible. However as an important finding is that there are no differences between doctors and nurses in predicting CIN2+ I recommend that a power analysis should be done post hoc, in order to avoid type II error. This is important as number of CIN2+ in the material is limited and somewhat lower than expected (n=39, distribution between examiner category is not shown). Minor points A flowchart of the study could be beneficial for the reader. I also want to know the number of women who actually provided biopsies, without doing my own calculations. The aim of the study is to evaluate accuracy in detecting of cervical lesions. In a strict statistical sense accuracy (accurate observations e.g. true positives + true negatives, divided by all observations) is not presented. It is done in another way, by ROC-curves, but this should preferably warrant a comment Table 1 CIN3+ usually comprises both CIN3 and "worse". Change to ICC (invasive cervical cancer) In conclusion this paper is of great interest but need some

	clarifications and revisions.

REVIEWER	Peter Sykes University Of Otago Christchurch New Zealand I know the author but otherwise have no competing or conflicting interests
REVIEW RETURNED	10-Jun-2014

GENERAL COMMENTS	I think this paper addresses three very relevant and important questions for cervical screening activities in low resource settings and because of this i would strongly support publication. however the study design, although satisfactory as a preliminary hypothesis generating study is not adequate to fully address the hypotheses. As a result the manuscript requires modification to take account of these limitations This interesting paper examines the accuracy of colposcopy using the swede score in a relatively low resource setting and compares the performance of nurses, doctors and two different instruments. The author in the discussion, then explores the idea of colposcopy by nurses using the swede score and the gynocular for primary screening or for treatment triage (dIagnosis) following VIA in a low resource setting. The hypotheses explored are thus; Nurse led colposcopy using the swede score has a similar performance (no significant difference) to Dr performed colposcopy. The gynocular has similar performance to a colposcope. Colposcopy by swede score is a suitable tool for screening and or diagnosis in a low resource setting. As such the design is complex and the results difficult to interpret Colposcopy was performed by both nurse and doctor using both gynocular and the colposcope on the same patient in an alternating fashion involving 6 cervical visualisation by both doctor and nurse on each patient. Colposcopy using the green light was performed by both nurse and doctor then inspection following wash with acetic acid by both nurse and doctor, the procedure was performed using both instruments the order chosen according to block randomisation. Swede score was then correlated with biopsy result, but biopst was only performed on women with an abnormal swede score, unfortunately the cut off score was 1 for nurses but 4 for doctors. There are several significant concerns regarding the methodology, It is hard to imagine in the circumstances that the nurse and doctor were blinded to each others colposcopic assesments, if not blinded this could well influence the correlation between practitioners. If blinded the steps taken to ensure blinding need to be described. Simmilarly the colposcopic examination was carried out using both instruments by the same colposcopist on all patients, it is very unlikely that scoring would be significantly different in these circumstances. Interpretation of the results needs to be very cautios, in these circumstances therefore there was no obvious difference between doctors or nurses or gynocular and colposcope although these

	results could not be generalised to other circumstances Afurther significant problem was the different threshold of biopsy between doctor and nurse this will certainly effect apparent performance differentially between doctors and nurses. It is not clear to me as to whether the study is adequately powered to detect a difference in performance between the two arms , there are a relatively small number of abnormal results no verification of normal and wide confidence intervals, however performance appeared very similar so this is of lesser concern. A formal adequately powered randomised trial, with adequate blinding or where one examination was carried out by one practitioner using a single instrument in each arm would be necessary to confirm the hypotheses above. The study was not designed to determine the utility of the swede score as a diagnostic or screening tool, there was no verification of normality in the patients with swede scores of less than 1 or less than 4 by doctors and the ROC curves do not indicate a high level of accuracy for the test for CIN2+ The positive predictive value of colposcopy with a cut off of 1 was 86% which is not substantially different from a cut off of 5 (87%) although this does not preclude the use of this strategy for screening or biopsy in a low resource setting , further studies including verification of normal by biopsy or hpv testing in at least a sample of those with swede scores of 0 and biopsy of all patients with swede score 1 or greater would be appropriate to explore this hypothesis further. In summary this paper generates but does not answer some very important questions, modification of the manuscript is suggested in order to account for the limitations of the study.

REVIEWER	Peter Baade Cancer Council Queensland Australia
REVIEW RETURNED	10-Jul-2014

GENERAL COMMENTS	This paper reports on a comparison of screening methods implemented by doctors and nurses for cervical cancer. This has clear implications for reducing the burden of cervical cancer in low resource settings, and any effective strategies that enable appropriately trained nurses to conduct screening in the absence of doctors should be encouraged. As outlined below, I did have some concerns regarding the use of these results as evidence to support the similarity of the different methods. An inclusion criteria (Line 181) was being VIA positive. This would increase the likelihood that the resulting colposcopy was also positive. However, in low resource settings (line 145) the colposcopy may be used as the primary screening tool, so the real-world prevalence of abnormal lesions would be much less in the 'real world' situation than in this study, with resulting implications for diagnostic accuracy. Line 157-160 – consider revising the paragraph for clarity. Are there 2 x 2 groups? Nurses/doctors and stationary/Gynocular? Line 262-263. Please clarify how you expect 2.5% of naïve women

	to have a positive biopsy result, when earlier (Line 170-171) you mentioned that 2.3% of women are screened with VIA and of those, 5% are VIA positive. Would this mean that 5% of the 2.3% are VIA positive? Line 261-265. How was “sufficient precision” defined? Given that the ideal outcome for this study is a null result (ie. no difference between nurses and doctors), it is important that evidence be provided that the study cohort is of sufficient size to enable adequate statistical power to detect a difference. More details are required here. Lines 271 and 275. No biopsy was taken for about 70% of women, This also has implications for the adequacy of the cohort size, because only 30% of the original sample have information about biopsy results, which are then used as the gold standard (line 222). Lines 284-290. The sensitivity and specificity scores are relatively low, considering that 50% is equivalent to tossing a coin. Lines 313-314. It is likely that the cross over design would have increased the similarity between the two screening procedures, because the information from one method would have influenced the perceptions of the other method. This would have the effect on favouring the null result (ie. no differences between the methods). Further information needs to be provided why the authors consider the risk nondifferential. Minor issues: English expression and grammar throughout the manuscript needs proof reading Abstract requires prior knowledge of terms. The VIA and Swede Score method need a brief explanation. Line 75 – suggest “no difference” be replaced with “no evidence of a significant difference” Line 127 – are these incidence data global estimates”? Line 146 – it would be useful to provide more background of the Swede score systematic colposcopy system. How widely is this used? How does it differ from other methods? Line 274. 303/528 (Table 1)=57.4%, not 58.2%

REVIEWER	Wojciech Rokita Department of Obstetrics and Gynecology Province Hospital Kielce Poland
REVIEW RETURNED	15-Jul-2014

- The reviewer completed the checklist but made no further comments

VERSION 1 – AUTHOR RESPONSE

Reviewer1

1:1 A quite crucial point in this study is blinding. Obviously there is no blinding between the instruments. This was not possible, but there is a bias introduced when the examiner examines the same women twice in a row with two different instruments. The conclusion that is drawn from the first examination of course can influence the second. The block randomisation prevents this favouring one instrument over the other, but as concordance is presented the procedure will bias the results. The ideal design might be that the same examination will see the woman a second time at a randomly selected moment, being assigned the colposcope not used the first time and being blinded to her identity. This is probably practically impossible, but the limitation should be discussed.

Revised
We have now accordingly adjusted the discussion.
1:2 Unfortunately there is no mentioning of blinding between nurses and doctors. It appears that both were present at the same time. Were there any provisions against communicating the findings between the examiners? A study nurse "immediately recorded" the scores. Was the communication with this nurse oral or in writing? Could the other examiner present overhear it? This should be stated to assess the validity of the results.
Revised
The question has been addressed in revised in the Methods section.
Thank you so much for this comment!
1:3 To assess the vessel patterns green filter without acetic acid was used. This is not in accordance with the original Swede score studies where assessment of mosaic pattern, punctuation, atypical vessels or absence of vessels was made after acetic acid application. This should be noted and discussed.
Revised
Revised as suggested. We appreciate this comment and have clarified the method section.
1:4 No basic studies have been done so far testing Swede score on postmenopausal women or breastfeeding women with atrophy of the vaginal mucosa. The actual range of age in this study is not presented, but the data in table 1 indicate that older women participated. I recommend that results from women in postmenopausal age as well as breastfeeding women, if these can be identified, should be analysed separately.
Revised
We thank this Reviewer for this comment. Data was not collected on breastfeeding and only one woman over 50 had a biopsy.
Revised in discussion.
1:5 Sensitivity and specificity are stated in Results and tables and I assume
this is related to find CIN2 or higher degrees of dysplasia or cancer (CIN2+) However this should be stated. The number of observations should be included in tables 2 and 3. According to the manuscript Swedescore rated by doctors ≥ 4 are included in the study, but level 4 is missing in table 3.
Revised
Thank very much for this comment. The comments have been addressed accordingly in the results and tables.
1:6 However as an important finding is that there are no differences between doctors and nurses in predicting CIN2+ I recommend that a power analysis should be done post hoc, in order to avoid type II error. This is important as number of CIN2+ in the material is limited and somewhat lower than expected (n=39, distribution between examiner category is not shown).

Revised
Thank you very much for this comment. A post hoc power analysis has been included in the

manuscript.
1:7 A flowchart of the study could be beneficial for the reader. I also want to know the number of women who actually provided biopsies, without doing my own calculations.
Revised
We clarified the number of biopsies in the manuscript in the result section. However, we did not include a flowchart as we have several tables and figures already. We clarified the study design in the method section. Thank you very much for this comment!
1:8 The aim of the study is to evaluate accuracy in detecting of cervical lesions. In a strict statistical sense accuracy (accurate observations e.g. true positives + true negatives, divided by all observations) is not presented. It is done in another way, by ROC-curves, but this should preferably warrant a comment.

Not revised
We acknowledge that there are other statistical methods that could have been used in order to evaluate accuracy. We have chosen to present sensitivity and specificity in ROC curves since it reflects both the impact on true positive and true negatives, which the measure of accurate observation don't.
1:9 Table 1
CIN3+ usually comprises both CIN3 and "worse". Change to ICC (invasive cervical cancer)
Revised
Table 1 and the text has been revised. Thank you for this valuable comment and all other comments that improved our manuscript!

Answers to comments and suggestions of Reviewer 2 Peter Sykes

We would like to thank Reviewer 2 very much for all his valuable comments that helped to improve our manuscript.

Reviewer2
2:1 I think this paper addresses three very relevant and important questions for cervical screening activities in low resource settings and because of this i would strongly support publication. however the study design, although satisfactory as a preliminary hypothesis generating study is not adequate to fully adress the hypotheses. As a result the manuscript requires modification to take account of these limitations

Revised
We have now further discussed the limitations of the study design in the discussion. Thank you for this valuable comment that improved our manuscript!
2:2 This interesting paper examines the accuracy of colposcopy using the swede score in a relatively low resource setting and compares the performance of nurses, doctors and two different instruments. The author in the discussion, then explores the idea of colposcopy by nurses using the swede score and the gynocular for primary screening or for treatment triage (dIagnosis) following VIA in a low resource setting.
The hypotheses explored are thus;
Nurse led colposcopy using the swede score has a similar performance (no significant difference) to Dr performed colposcopy.
The gynocular has similar performance to a colposcope.
Colposcopy by swede score is a suitable tool for screening and or diagnosis in a low resource setting.
Revised
The hypothesis has been revised in the introduction and in the discussion. Thank you for this comment
2:3 As such the design is complex and the results difficult to interpret

Colposcopy was performed by both nurse and doctor using both gynocular and the colposcope on the same patient in an alternating fashion involving 6 cervical visualisation by both doctor and nurse on each patient.

Colposcopy using the green light was performed by both nurse and doctor then inspection following wash with acetic acid by both nurse and doctor, the procedure was performed using both instruments the order chosen according to block randomisation.

Swede score was then correlated with biopsy result, but biopst was only performed on women with an abnormal swede score, unfortunately the cut off score was 1 for nurses but 4 for doctors.
There are several significant concerns regarding the methodology,
It is hard to imagine in the circumstances that the nurse and doctor were blinded to each others colposcopic assesments, if not blinded this could well influence the correlation between practitioners.
If blinded the steps taken to ensure blinding need to be described.

Revised
We have further described the methology and blinding in the methology section. We appreciate the comments. Thanks!
2:4 Simmilarly the colposcopic examination was carried out using both instruments by the same colposcopist on all patients, it is very unlikely that scoring would be significantly different in these circumstances.
Interpretation of the results needs to be very cautios, in these circumstances therefore there was no obvious difference between doctors or nurses or gynocular and colposcope although these results could not be generalised to other circumstances
Revised
The comments have been addressed accordingly in the discussion. Thank you for the comment!
2:5 A further significant problem was the different threshold of biopsy between doctor and nurse this will certainly effect apparent performance differentially between doctors and nurses.
It is not clear to me as to whether the study is adequately powered to detect a difference in performance between the two arms, there are a relatively small number of abnormal results no verification of normal and wide confidence intervals, however performance appeared very similar so this is of lesser concern.

Revised
We thank the reviewer 2 for this comment and has further clarified the threshold for biopsy in the discussion and the data in the figures and tables, only comparing Swede score 4 and above.
Revised
2:6 A formal adequately powered randomised trial, with adequate blinding or where one examination was carried out by one practitioner using a single instrument in each arm would be necessary to confirm the hypotheses above.

Revised
Thank you very much for this comment. Please find our comments in the methods section and the discussion.
2:7 The study was not designed to determine the utility of the swede score as a diagnostic or screening tool, there was no verification of normality in the patients with swede scores of less than 1 or less than 4 by doctors and the ROC curves do not indicate a high level of accuracy for the test for CIN2+

The positive predictive value of colposcopy with a cut off of 1 was 86% which is not substantially different from a cut off of 5 (87%) although this does not preclude the use of this strategy for

screening or biopsy in a low resource setting , further studies including verification of normal by biopsy or hpv testing in at least a sample of those with swede scores of 0 and biopsy of all patients with swede score 1 or greater would be appropriate to explore this hypothesis further.

Revised
The Swede score has been previously validated in several high resource settings and the threshold suggested for biopsy is a Swede score of 6. As we worked in a low resource setting we lowered the threshold for biopsy when a Swede score of 4, and have further clarified this in the manuscript, tables and figures. Also, in Ngonzi et al 350 patients were screened with VIA, 69 of them VIA positive. These women all had biopsy regardless of Swede score.
Thus, we did not aim to again evaluate the accuracy of the Swede score itself, rather how nurses compare to doctors when performing Swede score. Please find our revised discussion and power calculation. We very much appreciate your comments and they help us improving our manuscript!

Answers to comments and suggestions of Reviewer 3 Peter Baade.

We would like to thank Reviewer 3 and we so much appreciate all his valuable comments. The comments improved our manuscript. Thank you for all your efforts!

Reviewer3
3:1
An inclusion criteria (Line 181) was being VIA positive. This would increase the likelihood that the resulting colposcopy was also positive. However, in low resource settings (line 145) the colposcopy may be used as the primary screening tool, so the real-world prevalence of abnormal lesions would be much less in the 'real world' situation than in this study, with resulting implications for diagnostic accuracy.

Revised
In low resource settings, VIA screening is the most common screening method and screening colposcopy is not often used due to the technical limitations and costs of stationary colposcopes. This study included both VIA positive women (528) and women who came for screening (404). In the method section we described how both VIA positive women were included and women who were screening naïve, thus including both populations. Interestingly, few VIA positive women and screening naïve women were found to have cervical lesions. This result has implications for the diagnostic accuracy of VIA. We appreciate this comment! Not revised
3:2 Line 157-160 – consider revising the paragraph for clarity. Are there 2 x 2 groups? Nurses/doctors and stationary/Gynocular?
Revised
The question has been addressed in revised the method section. Thank you very much for this comment.
3:3 Line 262-263. Please clarify how you expect 2.5% of naïve women to have a positive biopsy result, when earlier (Line 170-171) you mentioned that 2.3% of women are screened with VIA and of those, 5% are VIA positive. Would this mean that 5% of the 2.3% are VIA positive?
Revised
Thank you for this comment. In Bangladesh, only opportunistic cervical screening data for VIA, HPV and cytology is available. Our Principal Investigator prof Ashrafunessa have published data from 2008-2010 where the rate of CIN in screening naïve women screened with cytology, ranged between 1.2%-3.3%. (Nessa et al Asian Pac J Cancer Prev. 2013;14(12):7607-11, Begum et al Mymensingh Med J. 2012 Jan;21(1):145-50. We therefore chose to assume a value "in between" of 2.5 % for positive biopsy result for the sample size estimation for screening naïve women. In VIA positive women and cytology positive women 7.7 % of them had CIN lesions in biopsy. We therefore chose the value of 7.5 % for VIA positive women in the sample size estimation.

VIA has a low sensitivity and specificity. Actually few women (around 7.5%) of the VIA positive women have CIN in biopsy from a colposcopy examination.

3:4 Line 261-265. How was “sufficient precision” defined? Given that the ideal outcome for this study is a null result (ie. no difference between nurses and doctors), it is important that evidence be provided that the study cohort is of sufficient size to enable adequate statistical power to detect a difference. More details are required here.

Revised
Please find our revision in the method section. We appreciate this comment and thank reviewer 2!
3:5 Lines 271 and 275. No biopsy was taken for about 70% of women, This also has implications for the adequacy of the cohort size, because only 30% of the original sample have information about biopsy results, which are then used as the gold standard (line 222).
Revised
Thank you for this comment! Swede score has been validated previously in Sweden and UK with a Swede score of 6 as as suggested threshold for biopsy . Swede score has also been used in previous studies in low resource settings. Please find our comments in the method and discussion.
3:6 Lines 284-290. The sensitivity and specificity scores are relatively low, considering that 50% is equivalent to tossing a coin.
Not revised
A high Swede score has a low sensitivity but a very high specificity. A low Swede score has a high sensitivity but a low sensitivity (Strander et al and Bowring et al). Thank you very much for this comment!
3:7 Lines 313-314. It is likely that the cross over design would have increased the similarity between the two screening procedures, because the information from one method would have influenced the perceptions of the other method. This would have the effect on favouring the null result (ie. no differences between the methods). Further information needs to be provided why the authors consider the risk nondifferential.

Revised
Thank you for this comment. We have revised our manuscript accordingly in the discussion.
3:8 Minor issues:
English expression and grammar throughout the manuscript needs proof reading

We had the manuscript proof read by a native English speaker. Thank you for this comment! Revised
3:9 Abstract requires prior knowledge of terms. The VIA and Swede Score method need a brief explanation.
Revised
Abstract revised. Thank you for this comment!
3:10 Line 75 – suggest “no difference” be replaced with “no evidence of a significant difference”

We appreciate this comment and have revised accordingly.
3:11
Line 127 – are these incidence data global estimates”?
Revised
Thank you for your comment. We have revised the introduction accordingly.
3:12 Line 146 – it would be useful to provide more background of the Swede score systematic colposcopy system. How widely is this used? How does it differ from other methods?

Revised as suggest in the introduction. Thank you very much for this comment!

3:13
Line 274. 303/528 (Table 1)=57.4%, not 58.2%
Revised
Revised as suggested in the result section.
We deeply appreciate all the efforts and comments of reviewer 3 that help us to improve our manuscript.

VERSION 2 – REVIEW

REVIEWER	Peter Sykes University Of Otago Christchurch New Zealand
REVIEW RETURNED	14-Sep-2014

GENERAL COMMENTS	This is an important study comparing the performance of colposcopy by nurses and doctors using a colposcope and the gynocular in a low resource setting. The crossover design limits the validity of the comparison of the colposcope and gynocular. The study is not designed to assess the sensitivity and sprcificity of the swede score as normality was not confirmed in those with a low swede score. This study is underpowered to show a significant difference in the performance eof either colposcope or colposcopist and the different threshold for biopsy also introduces a bias. However good correlations are seen in swede score between operators and colposcopes. the conclusion of the study can only be that in this setting sweede score as determined by doctor or nurse using the coposcope or gynocular are similar and offered similar positive predictive value for CIN2 +. The possible learning curve is worthy of mention and the implication that further study of nurse colposcopy using gynocular is warrented is important and justified. The authors need to be careful not to use the results of other studies to overinterpret the results of this. If certain sweede scores are recommended for clinical indications these need to be clearly justified. There is room to simplify the discussion s and reduce the number of figures, the gynocular is described elsewhere and could be referenced. In specific comment I don't believe the authors can use the words "highly accurate" in the abstract conclusion. The weaknesses of the study need to be further clarified and the paper needs further proof reading for a number of minor English grammatical errors provided these matters are addressed I would be happy to support publication of the paper without further review.

VERSION 3 – REVIEW

REVIEWER	Peter Sykes University Of Otago Christchurch New Zealand
REVIEW RETURNED	06-Oct-2014

GENERAL COMMENTS	This is much improved the one issue that continues to concern me is that the sensitivity of a swede score of 6 or greater for cin2+ is a little over 50% and most women were not biopsied so this may infact be an overestimate this is less than previously published this matter is not addressed