Received: 21 October 2019 | Revised: 29 March 2020 | Accepted: 4 May 2020

DOI: 10.1111/aogs.13895

ORIGINAL RESEARCH ARTICLE

AOGS Acta Obstetricia et Gynecologica Scandinavica

Second trimester cervical length measurements with transvaginal ultrasound: A prospective observational agreement and reliability study

Pihla Kuusela[1,2] | Ulla-Britt Wennerholm[1,3] | Helena Fadl[4] | Jan Wesström[5,6] | Peter Lindgren[7] | Henrik Hagberg[1,3] | Bo Jacobsson[3,8] | Lil Valentin[9,10]

[1]Center of Perinatal Medicine and Health, Institute of Clinical Sciences, Sahlgrenska Academy, Gothenburg, Sweden

[2]Department of Obstetrics and Gynecology, Södra Älvsborg Hospital, Borås, Sweden

[3]Department of Obstetrics and Gynecology, Region Vastra Gotaland, Sahlgrenska University Hospital, Gothenburg, Sweden

[4]Department of Obstetrics and Gynecology, Faculty of Medicine and Health, Örebro University, Örebro, Sweden

[5]Center for Clinical Research Dalarna, Falun Hospital, Falun, Sweden

[6]Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden

[7]Center for Fetal Medicine, Karolinska University Hospital, Stockholm, Sweden

[8]Department of Genetics and Bioinformatics, Domain of Health Data and Digitalization, Institute of Public Health, Oslo, Norway

[9]Department of Medical Sciences Malmö, Lund University, Lund, Sweden

[10]Department of Obstetrics and Gynecology, Skåne University Hospital, Malmö, Sweden

Correspondence
Lil Valentin, Department of Obstetrics and Gynecology, Skåne University Hospital, Jan Waldenströms gata 47, 20502 Malmö, Sweden.
Email: lil.valentin@med.lu.se

Funding information
The study was funded by Vetenskapsrådet (Swedish Research Council, Dnr 2014-06998), Forskning och Utbildning (FoU) Södra Älvsborg and from the Swedish state

Abstract

Introduction: Universal screening for preterm delivery by adding transvaginal ultrasound measurement of cervical length to routine second trimester ultrasound has been proposed. The aim is to estimate inter- and intraobserver agreement and reliability of second trimester transvaginal ultrasound measurements of cervical length performed by specially trained midwife sonographers.

Material and methods: This is a prospective reliability and agreement study performed in seven Swedish ultrasound centers. In total, 18 midwife sonographers specially trained to perform ultrasound measurements of cervical length and 286 women in the second trimester were included. In each center, two midwife sonographers measured cervical length a few minutes apart in the same woman, the number of women examined per examiner pair varying between 24 and 30 (LIVE study). Sixteen midwife sonographers measured cervical length twice ≥2 months apart on 93 video clips (CLIPS study). The main outcome measures were mean difference, limits of agreement, intraclass correlation coefficient, intra-individual standard deviation, repeatability, Cohen's kappa and Fleiss kappa.

Results: The limits of agreement and intraclass correlation coefficient of the best examiner pair in the LIVE study were −4.06 to 4.72 mm and 0.91, and those of the poorest were −11.11 to 11.39 mm and 0.31. In the CLIPS study, median (range) intra-individual standard deviation was 2.14 mm (1.40-3.46), repeatability 5.93 mm (3.88-9.58), intraclass correlation coefficient 0.84 (0.66-0.94). Median (range) interobserver agreement for cervical length ≤25 mm in the CLIPS study was 94.6% (84.9%-98.9%) and Cohen's kappa 0.56 (0.12-0.92), median (range) intraobserver agreement was 95.2% (87.1%-98.9%) and Cohen's kappa 0.68 (0.27-0.93).

Conclusions: Agreement and reliability of cervical length measurements differed substantially between examiner pairs and examiners. If cervical length measurements

Abbreviations: CI, confidence interval; GW, gestational week; ICC, intraclass correlation coefficient; IISD, intra-individual standard deviation.

under the agreement between the Swedish government and the county councils, the ALF-agreement (ALFGBG-136431, ALFGBG-426411).

are used to guide management there is potential for both over- and under-treatment. Uniform training and rigorous supervision and quality control are advised.

KEYWORDS

cervical length, cervical length measurement, data accuracy, inter-observer variation, intraobserver variation, preterm delivery, quality control, reproducibility of results, second trimester of pregnancy, transvaginal ultrasound

1 | INTRODUCTION

Preterm birth is a major cause of death before 5 years of age and of long-term morbidity from infancy to adulthood.[1,2] A sonographically short cervix in the second trimester is a risk factor for spontaneous preterm delivery[3] and universal screening for preterm delivery by adding transvaginal ultrasound measurement of cervical length to routine second trimester ultrasound has been proposed.[4] The definitions of short cervix vary, a common definition being ≤25 mm.[4] Cervical length measurements by ultrasound are considered to be easy to perform.[4] However, they are not without difficulties[5] and may be particularly problematic in early/mid second trimester when the isthmus is present.[6-8] In Sweden, all routine fetal ultrasound examinations are performed by specially trained midwives—midwife sonographers. These are certified by the Swedish Society of Obstetricians and Gynecologists to perform routine fetal ultrasound examinations after standardized theoretical and practical teaching and after having passed a theoretical and practical test. If screening for preterm delivery were to be introduced in Sweden, cervical length measurements would be performed by midwife sonographers. In this perspective, it is important to know the reproducibility and reliability of sonographic second trimester cervical length measurements carried out by midwife sonographers specially trained to perform these measurements.

Before the start of the current study we identified two studies that estimated inter- and intraobserver reproducibility of transvaginal ultrasound measurements of cervical length in the second trimester during live scanning[9,10] (search strategy in Appendix S1). Both were single-center studies and included few examiners (n = 4). Therefore, the generalizability of the results may be questioned. In everyday clinical practice, examiners with different levels of experience, skill and care perform ultrasound measurements of cervical length.

The aim of this study is to estimate inter- and intraobserver agreement and reliability of transvaginal ultrasound measurements of cervical length in the second trimester performed in different centers by a large number of midwife sonographers specially trained to measure cervical length.

2 | MATERIAL AND METHODS

This reproducibility study forms part of the CERVIX study, a prospective observational Swedish multicenter study performed between May 2014 and June 2017 (ISRCTN (http://www.isrctn.com/ISRCTN18093885). The aim of the CERVIX study is to estimate the sensitivity and specificity with regard to preterm delivery of cervical length as measured with ultrasound in the second trimester. Seven ultrasound centers participated. Asymptomatic women with a live singleton pregnancy attending a routine fetal ultrasound examination at 18^{+0} to 20^{+6} gestational weeks (GWs) were eligible. Gestational age was estimated on the basis of ultrasound measurement of the fetal biparietal diameter[11,12] as recommended in Swedish guidelines (https://www.sfog.se/media/336451/fetometri.pdf). Women who gave written informed consent underwent transvaginal ultrasound measurement of cervical length at 18^{+0} to 20^{+6} GWs and a repeat measurement at 21^{+0} to 23^{+6} GWs. Medical staff and the women themselves were blinded to the results, which were not used for clinical management.

The reproducibility study was performed between January 2016 and June 2017 and consists of the LIVE study and the CLIPS study. The main purpose of the LIVE study is to estimate interobserver agreement and that of the CLIPS study to estimate intraobserver measurement error and repeatability of cervical length measurements performed between 18^{+0} and 23^{+6} GWs.

All ultrasound examiners participating in the reproducibility study also examined women in the CERVIX study. The examiners were midwife sonographers with several years of ultrasound experience. They perform 400-2100 routine fetal ultrasound examinations per year (most of them approximately 1000 per year). All were certified to examine women in the CERVIX study after theoretical and practical training, that is, local hands-on training by a physician or midwife already certified and two theoretical lectures (one from The Fetal Medicine Foundation, https://fetalmedicine.org/cervical-assessment-1 and one created by the steering committee of the CERVIX study). A midwife sonographer was certified if all five members of

Key message

Universal screening for preterm delivery by transvaginal ultrasound measurement of cervical length has been proposed despite the reliability of such measurements being largely unknown. We found the reliability of cervical length measurements to differ substantially between examiners. Uniform training and careful supervision are recommended.

the quality control committee of the CERVIX study agreed that her ultrasound images from five consecutive women (three images per woman, in all 15 images) examined at 18-23 GWs fulfilled our quality criteria. The quality control committee consisted of LV, BJ and JW, who were head of their respective ultrasound unit during the study period, and two midwife sonographers certified to perform cervical length measurements in the CERVIX study. Our quality criteria were:

- the cervix occupies at least 75% of the screen,
- the anterior and posterior lip of the cervix are of equal thickness,
- the full length of the endocervical canal is clearly seen,
- the inner and outer cervical os are clearly seen as well as the virtual inner os if isthmus is present (isthmus is the lowest part of the uterine corpus that develops into the lower uterine segment as pregnancy progresses),
- calipers are positioned correctly at the internal and external os and at the virtual inner os if isthmus is present (Figure 1).

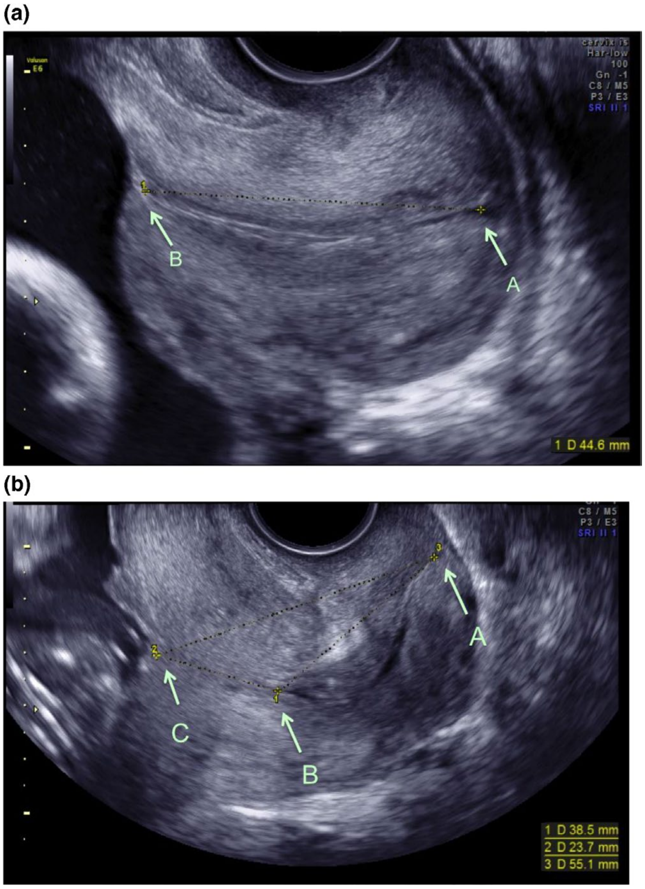


FIGURE 1 Measurement of cervical length when isthmus is absent (a) or present (b). A denotes the external os, B denotes the internal os. C (which we call the "virtual inner os") is the innermost end of the juxtaposed anterior and posterior isthmus (isthmus is the lowest part of the uterine corpus that develops into the lower uterine segment as pregnancy progresses). Measurements are taken as a straight line from A to B (endocervical length), B to C (isthmus length) and A to C [Color figure can be viewed at wileyonlinelibrary.com]

Twenty-five midwife sonographers were certified to perform cervical length measurements. After certification, quality controls were performed four times a year. Two quality checks were pre-planned but the midwives were not notified in advance of the other two. For the pre-planned quality controls, all certified midwives selected three ultrasound images of the cervix, considered by themselves to fulfil our quality criteria, and submitted them to the quality control committee via a web-based system (MedSciNet AB, Stockholm, Sweden, www.medscinet.com). The images were evaluated using the web-based software and had to fulfil all five quality criteria. If all three images were not approved, new images needed to be submitted within 1 month. After three subsequent failed quality checks, the midwife was no longer allowed to examine women in the CERVIX study. For quality controls carried out without prenotification, ultrasound images of the cervix from five consecutive women (three images per woman) examined during 1 week, specified by the quality control committee, were selected retrospectively by each midwife and submitted to the quality control committee for evaluation. The committee gave feedback and suggestions for improvement when needed. Educational material, documents and images describing our quality criteria and measurement technique were available for consultation throughout the study on the website of the CERVIX study. To ensure uniform measurement technique, during the course of the study, two midwife sonographers from each center paid a 2-day visit to the coordinating center at Skåne University Hospital to follow the midwife sonographers at that center (one of them being a member of the quality control committee) at which time they performed their cervical length measurements and discussed problems and pitfalls.

The cervical length measurements in the LIVE and CLIPS study were performed in the same manner as in the CERVIX study. The women were examined in the lithotomy position in a gynecological chair with an empty urinary bladder. The transvaginal probe was introduced into the vagina. A sagittal view of the cervix was obtained. The length of the endocervical canal (distance A-B) was measured as a straight line from the external to the internal cervical os. If the isthmus was present, three distances were measured: the endocervical length (distance A-B), the isthmus length (distance B-C) and the distance A to C (Figure 1). Funneling was not recorded and fundal or suprapubic pressure was not applied. Three measurements of each distance were taken during at least 3 min, each measurement being taken on a new image. All measurements were recorded in our web-based electronic case record form (MedSciNet AB, www.medscinet.com). It was obligatory to store still images of all measurements electronically. The ultrasound systems used were a GE Healthcare Voluson E8 Expert or E6 with a 5-9 MHz vaginal transducer (GE Corporate).

For the LIVE study, the women underwent ultrasound examination of the cervix by two midwife sonographers working in the same center in the same scanning session. One pair of midwives from each center—selected on the basis of their availability for the study (eg part-time or full-time employment)—participated in the LIVE study. The examiners took turns to take the measurements first or secondly, with the shortest possible interval (a few minutes) between the two examinations. They were blinded to each other's results: only the midwife performing the examination was present in the examination room. No ultrasound images of the cervix were left on the ultrasound screen for the second examiner to see. The results of the midwife who examined first were used both in the CERVIX study and in the LIVE study, the second examiner's results were used only in the LIVE study. A woman could participate only once in the LIVE study.

For the purpose of the CLIPS study, video clips were collected from the volunteers participating in the CERVIX study and were evaluated by midwife sonographers participating as examiners in the CERVIX study. Video clips with a duration of 8-10 seconds of consecutive ultrasound examinations of the cervix were collected for the purpose of the CLIPS study by the midwife sonographer who was a member of the quality control committee when she measured the cervix of women participating in the CERVIX study (her measurements were used in the CERVIX study). The video clips were distributed to the examiners in the CLIPS study. They were analyzed on an ultrasound system with measurements of cervical length being taken using the measurement function of the ultrasound machine. The same measurement technique as described above was used, but only one measurement per distance was taken. Each rater analyzed the clips twice ≥2 months apart and in a different order. All raters were blinded to the results of the other raters and to their own previous results. A woman could participate only once in the CLIPS study.

Formal sample size calculation was not performed. Sample size was based on availability of ultrasound examiners. We planned to let one examiner pair per center examine 30 women and to let all midwife sonographers certified to measure cervical length (except the one who collected the video clips for the CLIPS study) assess 100 video clips. We designed the study and prepared the manuscript following the Guidelines for reporting reliability and agreement studies.[13]

2.1 | Statistical analyses

Below is a short description of our statistical analysis; details are presented in Appendix S2.

We use the terminology recommended in Guidelines for reporting reliability and agreement studies: "Agreement" is the degree to which measurements are identical and "reliability" is the ability of a measurement or a categorical variable to differentiate between subjects.[13]

We assessed the relation between the interobserver differences and the magnitude of the measurement values by plotting the absolute interobserver differences against the mean measurement results of the two examiners in the same examiner pair,[14] and that between intraobserver differences and the magnitude of the measurement values by plotting the intra-individual standard deviation (IISD) against the mean of each rater's measurements (Bland-Altman plots).[15]

We express interobserver agreement for continuous measurements as mean difference and limits of agreement (95% of differences between future measurements by two examiners in a pair are expected to fall between these limits).[14] For assessment of systematic differences between two examiners we present the 95% confidence interval (CI) of the mean difference.

We describe the ability of observers to reproduce their own results ("intraobserver agreement") as measurement error (IISD) and repeatability (2.77 × IISD). The difference between a subject's measurement and the true value is expected to be less than 1.96 × IISD for 95% of observations.[16] The difference between two measurements on the same subject is expected to be less than the repeatability in 95% of pairs of observations.[16]

We express inter- and intraobserver reliability of continuous measurements as the intraclass correlation coefficient (ICC) with 95% CI.[17] ICC is the proportion of variance between examined individuals and the total variance. The higher the variance between examined individuals, the higher the ICC, in particular if the intra-individual variance is also small.[18]

We present inter- and intraobserver agreement with regard to the presence of isthmus and shortest endocervical length ≤25 mm as total percentage agreement, positive agreement and negative agreement.[19] We use Cohen's kappa and Fleiss kappa[20,21] as estimates of reliability. Some suggest that Cohen's kappa 0.81-1 indicates very good agreement, kappa 0.61-0.80 good agreement, kappa 0.41-0.60 moderate agreement, kappa 0.21-40 fair agreement and kappa ≤0.20 poor agreement.[22]

For all calculations we used the statistical software SAS System Version 9.4 (SAS Institute).

2.2 | Ethical approval

The CERVIX study including the reproducibility study was approved by the Regional Ethical Committee at the University of Gothenburg (Dnr 825-13 date of approval 11 November 2013, Dnr T053-14 date of approval 21 January 2014, Dnr T691-14 date of approval 19 September 2014, Dnr T972-15 date of approval 7 December 2015, Dnr T122-16 date of approval 25 February 2016, Dnr T896-17 date of approval 16 October 2017, Dnr T645-18 date of approval 2018-07-09, Dnr T878-18 date of approval 2018-10-11, Dnr T970-18 date of approval 1 November 2018). Clinical trial number: ISRCTN18093885.

3 | RESULTS

3.1 | LIVE study

Seven examiner pairs (14 midwife sonographers) participated in the LIVE study, one pair from each ultrasound center. Before participation in the LIVE study, two of the 14 midwife sonographers had performed cervical length measurements in <100 women, six in 100-299 women, four in 300-499 women and two in ≥500 women. Demographic details of the 198 pregnant participants are shown in

Table S1. Mean (SD) age was 31.5 (4.8) years, 85% of the participants were white, 42% were nulliparous, mean (SD) body mass index at the first antenatal visit was 25.5 (4.7) kg/m^2 (range 17.2-41.2). All cervix measurements were performed between 18^{+0} and 23^{+6} GWs.

Agreement and reliability with regard to presence of isthmus differed substantially between the examiner pairs (Table S2). Median total agreement was 93.3% (range 82.8%-96.4%). Negative agreement was good (median 95.8%, range 87.8%-98.2%) and superior to positive agreement (median 70.6%, range 0%-94.7%). Median Cohen's kappa was 0.69 (range 0.27-0.91). In 34 (17.2%) of the 198 women examined in the LIVE study, both examiners in an examiner pair agreed that the isthmus was present, and in 17 (8.6%) one of the examiners in a pair recorded the isthmus to be present.

A summary of the results for all seven examiner pairs for all 12 measurements—mean, minimum and maximum of endocervical length, distance A-C, isthmus length and isthmus length plus endocervical length—is shown in Table S3. There was no obvious trend for interobserver reliability to be better for any of the measurements. Below we present results for shortest endocervical length, which seems to be the measurement most often used in clinical practice and scientific studies.[4,23]

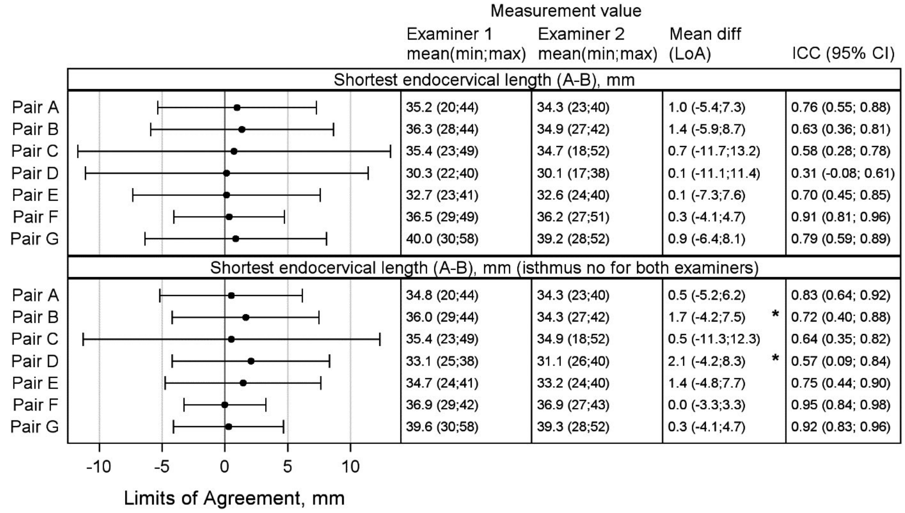
The interobserver differences in measurements of shortest endocervical length did not change with increasing measurement values. Forest plots showing the mean difference and limits of agreement of seven examiner pairs for measurements of shortest endocervical length and for shortest endocervical length when both examiners in a pair agreed on absence of isthmus are presented in Figure 2. We present a summary of the results with regard to shortest endocervical length for all seven examiner pairs in Table S3. For the examiner pair with the best interobserver agreement and reliability, the mean difference between the two examiners' measurements of shortest endocervical length was 0.33 mm (95% CI −.61 to 1.28), the limits of agreement were −4.06 to 4.72 mm and the ICC was 0.91. For those two pairs with poorest interobserver agreement and reliability, the mean differences for measurement of shortest endocervical length were 0.73 mm (95% CI −1.64 to 3.10) and 0.14 mm (95% CI −2.08 to 2.37), the limits of agreement were −11.70 to 13.17 mm and −11.11 to 11.39 mm, and the ICC 0.58 and 0.31. Agreement and reliability with regard to measurement of shortest endocervical length were better if both examiners in a pair agreed that the isthmus was absent (Figure 2, Table S3).

Because the shortest endocervical length ≤25 mm was recorded by at least one examiner in only four examiner pairs, agreement and reliability with regard to the shortest endocervical length ≤25 mm cannot be reliably estimated in the LIVE study (Table S4).

Measurement error differed between the 14 ultrasound examiners in the LIVE study (Table S5). For shortest endocervical length measurement error (IISD) was less than 1 mm for eight examiners, 1-1.5 mm for five examiners and 2.4 mm for one examiner.

3.2 | CLIPS study

Sixteen midwife sonographers, here called raters, participated in the CLIPS study; 12 of them also participated in the LIVE study. Sixteen raters yield 120 rater pairs. Nine of the 25 certified



	Measurement value Examiner 1 mean(min;max)	 Examiner 2 mean(min;max)	 Mean diff (LoA)	 ICC (95% CI)
Shortest endocervical length (A-B), mm				
Pair A	35.2 (20;44)	34.3 (23;40)	1.0 (-5.4;7.3)	0.76 (0.55; 0.88)
Pair B	36.3 (28;44)	34.9 (27;42)	1.4 (-5.9;8.7)	0.63 (0.36; 0.81)
Pair C	35.4 (23;49)	34.7 (18;52)	0.7 (-11.7;13.2)	0.58 (0.28; 0.78)
Pair D	30.3 (22;40)	30.1 (17;38)	0.1 (-11.1;11.4)	0.31 (-0.08; 0.61)
Pair E	32.7 (23;41)	32.6 (24;40)	0.1 (-7.3;7.6)	0.70 (0.45; 0.85)
Pair F	36.5 (29;49)	36.2 (27;51)	0.3 (-4.1;4.7)	0.91 (0.81; 0.96)
Pair G	40.0 (30;58)	39.2 (28;52)	0.9 (-6.4;8.1)	0.79 (0.59; 0.89)
Shortest endocervical length (A-B), mm (isthmus no for both examiners)				
Pair A	34.8 (20;44)	34.3 (23;40)	0.5 (-5.2;6.2)	0.83 (0.64; 0.92)
Pair B	36.0 (29;44)	34.3 (27;42)	1.7 (-4.2;7.5) *	0.72 (0.40; 0.88)
Pair C	35.4 (23;49)	34.9 (18;52)	0.5 (-11.3;12.3)	0.64 (0.35; 0.82)
Pair D	33.1 (25;38)	31.1 (26;40)	2.1 (-4.2;8.3) *	0.57 (0.09; 0.84)
Pair E	34.7 (24;41)	33.2 (24;40)	1.4 (-4.8;7.7)	0.75 (0.44; 0.90)
Pair F	36.9 (29;42)	36.9 (27;43)	0.0 (-3.3;3.3)	0.95 (0.84; 0.98)
Pair G	39.6 (30;58)	39.3 (28;52)	0.3 (-4.1;4.7)	0.92 (0.83; 0.96)

FIGURE 2 LIVE study. Forest plots showing mean difference (dot) and limits of agreements (LoA; lines) for seven examiner pairs for measurements of shortest endocervical length (A-B) and for shortest endocervical length when both examiners in a pair agreed on absence of isthmus. Each examiner pair consists of two midwives working in the same center. Measurements and differences are shown in mm, and all mean differences are shown as positive differences with LoA adjusted accordingly. Asterisks denote systematic differences between the two examiners. The measurement results and the intraclass correlation coefficient (ICC) with its 95% confidence interval (CI) for each pair are also shown. Examiner 1 is the examiner with the highest mean value of the studied variable. Examiner 2 is the examiner with the lowest mean value of the studied variable. We expect 95% of differences between future measurements by the two examiners in a pair to fall between the limits of agreement.[14] The ICC is considered to be a measure of reliability, that is, to reflect how well the measurements can discriminate between different individuals. ICC depends on the variance in the population studied. The higher the variance, the higher the ICC, in particular if the intra-individual variance ("measurement error") is also small[18]

midwife sonographers could not participate in the CLIPS study, one because she collected the video clips and eight because they changed workplace or were on long-term sick leave. Before participation in the CLIPS study, three of the 16 midwife sonographers had performed cervical length measurements in <100 women, three in 100-299 women, eight in 300-499 women and two in ≥500 women. Of 318 video clips, 100 were judged to be of very high quality by the midwife who collected the clips and were selected by her for use in the CLIPS study. Seven of the 100 clips were excluded because seven women contributed two clips, one from the examination at 18-20 GWs and another from the examination at 21-23 GWs. We used the former. Demographic details of the 93 pregnant participants in the CLIPS study are shown in Table S1. Mean (SD) age was 31.0 (4.1) years, 88% of the participants were white, 51% were nulliparous and mean (SD) body mass index at the first antenatal visit was 24.7 (4.3) kg/m^2 (range 18.2-42.3). All cervical length measurements were performed between 18^{+0} and 23^{+4} GWs.

Intraobserver agreement and reliability with regard to the presence of isthmus differed substantially between the 16 raters (Table S6). Median (range) total agreement was 89.2% (74.2%-94.6%), negative agreement 93.4% (80.3%-97.1%), positive agreement 70.7% (40.0%-90.2%) and Cohen's kappa 0.66 (0.36-0.87).

The IISD for endocervical length did not change with the mean of the measurement values. The 16 raters differed substantially with regard to their ability to reproduce their own results, see Figure 3, which shows Bland-Altman plots, IISD and ICC for each rater. The median (range) mean difference between two repeated measurements of endocervical length for the 16 raters was −0.15 mm (−1.48 to 1.27), IISD 2.14 mm (1.40-3.46), repeatability 5.93 mm (3.88-9.58) and ICC 0.84 (0.66-0.94) (Table S7). The 95% CIs for the individual ICCs are presented in Table S8 together with intraobserver measurement error and reliability for distance A-C, isthmus length and isthmus length plus endocervical length. Measurement error (IISD) tended to be smaller, repeatability better (smaller measurement errors) and ICC values to be higher for examinations performed at 21-23 GWs (n = 20) than for examinations performed at 18-20 GWs (n = 73) (Table S9).

Intraobserver agreement and reliability with regard to endocervical length ≤25 mm are shown in Table 1. The median (range) total agreement for the 16 raters was 95.2% (87.1%-98.9%), negative agreement 97.4% (93.3%-99.4%), positive agreement 69.7% (28.6%-93.3%) and Cohen's kappa 0.68 (0.27-0.93). Cohen's kappa was ≥0.60 (ie at least good) for 10/16 (62.5%) raters and ≥0.40 (ie at least fair) for 15/16 (93.8%) raters.

For estimation of interobserver agreement and reliability of the 120 rater pairs in the CLIPS study we used the results of the first analysis round. Interobserver agreement and reliability with regard to the presence of isthmus were median (range) total agreement 81.7% (59.1%-93.5%), negative agreement 88.9% (71.3%-96.5%), positive agreement 52.0% (22.2%-92.7%) and Cohen's kappa 0.42 (0.12-0.87). Fleiss kappa was 0.41 (95% CI .39-.43). The interobserver differences and limits of agreement for endocervical length varied widely between the examiner pairs and were similar to those in the LIVE study (Table S10). The ICC value was 0.67 (95% CI .60-.74). The limits of agreement tended to be narrower and the ICC value to be higher for examinations performed at 21-23 GWs (n = 20) than for examinations at 18-20 GWs (n = 73) (Table S11). Interobserver agreement and reliability with regard to endocervical length ≤25 mm were median (range) total agreement 94.6% (84.9%-98.9%), negative agreement 97.1% (92.3%-99.5%), positive agreement 58.8% (13.3%-92.3%) and Cohen's kappa 0.56 (0.12-0.92). Fleiss kappa was 0.53 (95% CI .51-.55).

4 | DISCUSSION

We found substantial differences between examiner pairs with regard to interobserver agreement and reliability of cervical length measurements, and substantial differences between examiners with regard to measurement error, repeatability and reliability.

Our study is the first to estimate reproducibility, repeatability and reliability of second trimester cervical length measurements that involve a large number of ultrasound examiners, examiners from different ultrasound centers and examiners that are sonographers, not doctors. Our results should be generalizable to examiners with similar training, supervision and quality control and should reflect reality better than those of single-center studies involving only a few examiners. The ultrasound experience and competence of the midwife sonographers who performed the cervical length measurements in our study are likely to be similar to those of certified sonographers who perform obstetric scans.

It may be seen as a limitation that we did not test intraobserver repeatability and reliability in our LIVE study. However, during live scanning, examiners cannot be blinded to their own results because measurements must be repeated within a short interval. This means that results calculated from measurements taken during live scanning do not truly reflect intraobserver repeatability. Instead we studied intraobserver agreement and reliability in the CLIPS study, the results of which are not completely generalizable to a live situation. Assessment of video clips eliminates variation due to image acquisition and physiological changes over time, and taking measurements on video clips is different from taking them during a live scan. Using video clips made it possible to estimate interobserver agreement for a very large number of examiner pairs (n = 120) and pairs consisting of examiners from different centers.

The shortest of three endocervical length measurements seems to be the measurement most often used in research and clinically.[3,10,23-28] However, in many studies it is unclear whether repeated measurements were taken and, if they were taken, which measurement result was used.[29-32] Only three studies describe how to measure cervical length when the isthmus is present.[6-8] We suspect that isthmus length is sometimes included in what is incorrectly judged to be endocervical length. Moreover, if the isthmus is present, the inner cervical os is difficult to define and measurements may be inaccurate. Our results show that interobserver agreement for endocervical length was better when the isthmus

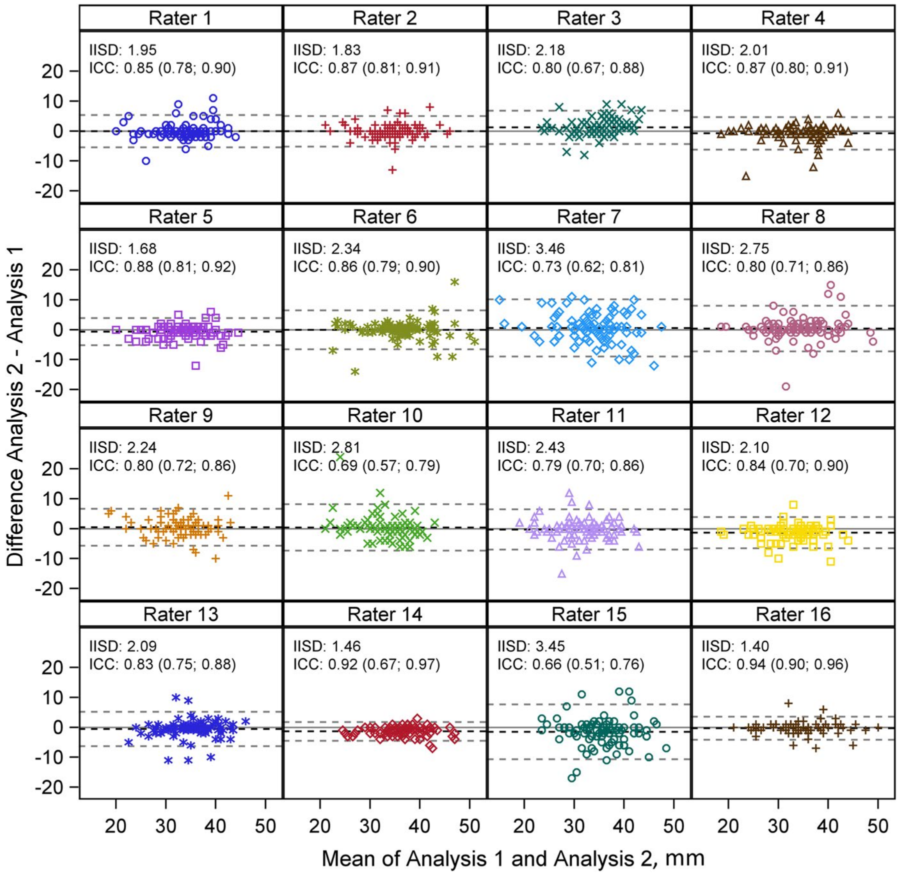


FIGURE 3 CLIPS study. Bland-Altman plots showing intraobserver differences (mm) between the second and first analysis of video clips for measurement of endocervical length (mm) for 16 individual raters. The mean of the two measurements is shown on the *x*-axis, and the difference between the second and first measurement is shown on the *y*-axis. The black-dotted horizontal line denotes the mean difference, the gray horizontal dotted lines denote the limits of agreement (within which 95% of the differences fall). The intra-individual SD (IISD) and the intraclass correlation coefficient (ICC) with its 95% confidence interval in brackets are also shown. The difference between a subject's measurement and the true value is expected to be <1.96 × IISD for 95% of the observations.[16] The difference between two measurements on the same subject (repeatability) is expected to be less than 2.77 × IISD in 95% of pairs of observations.[16] The ICC is considered to be a measure of reliability, that is, to reflect how well the measurement can discriminate between different individuals. It depends on the variance in the population studied. The higher the variance, the higher the ICC, in particular if the intra-individual variance ("measurement error") is also small[18] [Color figure can be viewed at wileyonlinelibrary.com]

was judged to be absent by both examiners in an examiner pair. The tendency for reproducibility, repeatability and reliability to be better for examinations performed at 21-23 GWs than at 18-20 GWs is likely to be explained by the isthmus being present less often at 21-23 GWs. Recognizing the isthmus and how to measure cervical length when the isthmus is present should be included in teaching cervical length measurements with ultrasound.

Many would probably—arbitrarily—find inter- and intraobserver differences in cervical length of up to 5 mm clinically acceptable, but differences ≥10 mm unacceptable. Whether differences of 6, 7, 8 or 9 mm are acceptable is debatable. For the observer pair in our LIVE study with the best interobserver agreement, one can expect 95% of any future differences between them to fall between −4.06 and 4.72 mm. This is similar to the interobserver agreement reported by Heath et al,[10] but the examiners in their study (who were probably doctors, their profession is not clearly described) were not blinded to each other's results and the mean of two measurements—not the shortest of three—was compared. For two examiner pairs in our LIVE study, the limits of agreement were approximately ±11 mm. The interobserver limits of agreement and ICC values reported in the study by França et al,[9] in which cervical length was measured by doctors, are similar to ours, but it is unclear which measurements they compared—possibly the first of three, not the shortest of three. Souka and Pilalis recently reported intraobserver agreement and reliability for one doctor (mean difference −0.5 mm, limits of agreement −3.5 to 2.5; ICC 0.98) and interobserver agreement and reliability for one pair of doctors (mean difference 1 mm, limits of agreement −4.7 to 6.7; ICC 0.93).[33] For the best rater in our CLIPS study, the difference between two measurements taken on the same subject is expected to be ≤3.9 mm in 95% of pairs of observations (repeatability 3.9 mm). The repeatability was ≤5 mm for five of 16 raters and 9.6 mm for the poorest rater. The substantial differences in results between examiner pairs and between examiners are likely to be explained by differences in skill and care of the examiners, in the stringency of local supervision, and by differences in local training.

Even though the relation between cervical length and preterm delivery is a continuum (the shorter the cervix, the higher the likelihood

TABLE 1 CLIPS study. Intraobserver agreement and reliability with regard to endocervical length ≤25 mm (Yes/No) for 16 raters

	Analysis 1 (A1) compared with analysis 2 (A2) with regard to endocervical length ≤25 mm				Agreement			Reliability
Unique rater number	Both No n (%)	Both Yes n (%)	A1 No, A2 Yes n (%)	A2 No, A1 Yes n (%)	Total agreement n (%) (95% CI)	Positive agreement[a] n/n (%)	Negative agreement[b] n/n (%)	Cohen's kappa (95% CI)
Rater 1	85 (91.4)	7 (7.5)	1 (1.1)	0 (0.0)	92 (98.9) (94.2-100.0)	14/15 (93.3)	170/171 (99.4)	0.93 (.79-1.00)
Rater 2	86 (92.5)	2 (2.2)	1 (1.1)	4 (4.3)	88 (94.6) (87.9-98.2)	4/9 (44.4)	172/177 (97.2)	0.42 (.01-.83)
Rater 3	85 (91.4)	3 (3.2)	1 (1.1)	4 (4.3)	88 (94.6) (87.9-98.2)	6/11 (54.5)	170/175 (97.1)	0.52 (.15-.88)
Rater 4	85 (91.4)	6 (6.5)	2 (2.2)	0 (0.0)	91 (97.8) (92.4-99.7)	12/14 (85.7)	170/172 (98.8)	0.85 (.64-1.00)
Rater 5	86 (92.5)	4 (4.3)	3 (3.2)	0 (0.0)	90 (96.8) (90.9-99.3)	8/11 (72.7)	172/175 (98.3)	0.71 (.40-1.00)
Rater 6	85 (91.4)	4 (4.3)	2 (2.2)	2 (2.2)	89 (95.7) (89.4-98.8)	8/12 (66.7)	170/174 (97.7)	0.64 (.32-.97)
Rater 7	76 (81.7)	5 (5.4)	2 (2.2)	9 (9.7)	81 (87.1) (79.6-93.9)	10/21 (47.6)	152/163 (93.3)	0.42 (.14-.69)
Rater 8	84 (90.3)	6 (6.5)	2 (2.2)	1 (1.1)	90 (96.8) (90.9-99.3)	12/15 (80.0)	168/171 (98.2)	0.78 (.54-1.00)
Rater 9	83 (89.2)	6 (6.5)	3 (3.2)	1 (1.1)	89 (95.7) (89.4-98.8)	12/16 (75.0)	166/170 (97.6)	0.73 (.47-.98)
Rater 10	83 (89.2)	5 (5.4)	1 (1.1)	4 (4.3)	88 (94.6) (87.9-98.2)	10/15 (66.7)	166/171 (97.1)	0.64 (.35-.93)
Rater 11	78 (83.9)	9 (9.7)	3 (3.2)	3 (3.2)	87 (93.5) (86.5-97.6)	18/24 (75.0)	156/162 (96.3)	0.71 (.50-.93)
Rater 12	79 (84.9)	7 (7.5)	5 (5.4)	0 (0.0)	86 (92.5) (87.6-98.2)	14/19 (73.7)	158/163 (96.9)	0.71 (.47-.95)
Rater 13	88 (94.6)	2 (2.2)	3 (3.2)	0 (0.0)	90 (96.8) (90.9-99.3)	4/7 (57.1)	176/179 (98.3)	0.56 (.12-1.00)
Rater 14	87 (93.5)	1 (1.1)	5 (5.4)	0 (0.0)	88 (94.6) (87.9-98.2)	2/7 (28.6)	174/179 (97.2)	0.27 (−.15 to .70)
Rater 15	85 (91.4)	3 (3.2)	3 (3.2)	2 (2.2)	88 (94.6) (87.9-98.2)	6/11 (54.5)	170/175 (97.1)	0.52 (.15-.89)
Rater 16	86 (92.5)	4 (4.3)	2 (2.2)	1 (1.1)	90 (96.8) (90.9-99.3)	8/11 (72.7)	172/175 (98.3)	0.71 (.40-1.00)
Distribution for all 16 raters								
Mean					95.09	65.52	97.44	0.63
Median (Min; Max)					95.2 (87.1-98.9)	69.7 (28.6-93.3)	97.4 (93.3-99.4)	0.68 (.27-.93)

Some suggest that Cohen's kappa 0.81-1 indicates very good agreement, kappa 0.61-0.80 good agreement, kappa 0.41-0.60 moderate agreement, kappa 0.21-40 fair agreement and kappa <0.20 poor agreement.[22]

A1, Analysis 1, that is, first Analysis of video clips; A2, Analysis 2, that is, second Analysis of video clips 2 months later.

[a] Positive agreement = (2 × Yes both Analyses)/(Yes Analysis 1 + Yes Analysis 2).[19]

[b] Negative agreement = (2 × No both Analyses)/(No Analysis 1 + No Analysis).[19]

of preterm delivery[3]), cervical length ≤25 mm is often used to identify pregnant women at high risk of preterm delivery.[25,27,31,32,34-37] Therefore, some might argue that precise measurements are not needed if the cervix can be reliably classified as ≤25 or >25 mm. On the other hand, some suggest that shortening of the cervix is a predictor of preterm delivery.[25,38,39] To detect changes, repeatability and interobserver agreement must be good. In our study, agreement that the cervix was >25 mm (negative agreement) was generally good and agreement that it was ≤25 mm (positive agreement) was poorer. This means that if cervical length is reported to be ≤25 mm, re-assessment may be wise, since a finding of a "short cervix" may instigate medical intervention. The importance of quality assessment and the need for programs to educate and certify sonographers before considering universal cervical length screening has been emphasized by others.[40,41] However, how educational programs and practical training are best organized is not known. As pointed out by Boelig et al,[40] more research is needed in this area. We speculate that uniform centralized education and training, rigorous post-training local supervision and post-training central quality control may be the way forward.

5 | CONCLUSION

Despite practical training, theoretical lectures, stringent criteria to obtain certification to perform cervical length measurements and four post-training annual quality controls, there were large differences between the ultrasound examiners in this study with regard to their ability to reproduce their own results and those of others. Intraobserver agreement with regard to cervical length ≤25 mm was good or very good for about 60% of the examiners and interobserver agreement was at least fair for 50% of the examiner pairs. If cervical length measurements with ultrasound are used clinically to guide management there is potential for both over- and under-treatment. Uniform training and rigorous post-training quality control are advised.

ACKNOWLEDGMENTS

Anders Pehrsson and Nils-Gunnar Pehrsson, the Statistical Consulting Group, Gothenburg, Sweden, performed the statistical analyses. Eva Hessman, librarian at Biomedical Libraries, University of Gothenburg, Sweden, helped with the literature search.

CONFLICT OF INTEREST

The authors have stated explicitly that there are no conflicts of interest in connection with this article.

ORCID

Pihla Kuusela https://orcid.org/0000-0001-9915-1201
Ulla-Britt Wennerholm https://orcid.org/0000-0003-2475-2226
Helena Fadl https://orcid.org/0000-0002-2691-7525
Jan Wesström https://orcid.org/0000-0003-1907-1071
Peter Lindgren https://orcid.org/0000-0001-5493-3402
Henrik Hagberg https://orcid.org/0000-0003-3962-1448
Bo Jacobsson https://orcid.org/0000-0001-5079-2374
Lil Valentin https://orcid.org/0000-0002-3830-6414

REFERENCES

1. Liu LI, Oza S, Hogan D, et al. Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. *Lancet*. 2016;388:3027-3035.
2. Saigal S, Doyle LW. An overview of mortality and sequelae of preterm birth from infancy to adulthood. *Lancet*. 2008;371:261-269.
3. Iams JD, Goldenberg RL, Meis PJ, et al. The length of the cervix and the risk of spontaneous premature delivery. National Institute of Child Health and Human Development Maternal Fetal Medicine Unit Network. *N Engl J Med*. 1996;334:567-572.
4. Campbell S. Prevention of spontaneous preterm birth: universal cervical length assessment and vaginal progesterone in women with a short cervix: time for action! *Am J Obstet Gynecol*. 2018;218:151-158.
5. Yost NP, Bloom SL, Twickler DM, Leveno KJ. Pitfalls in ultrasonic cervical length measurement for predicting preterm birth. *Obstet Gynecol*. 1999;93:510-516.
6. Greco E, Lange A, Ushakov F, Calvo JR, Nicolaides KH. Prediction of spontaneous preterm delivery from endocervical length at 11 to 13 weeks. *Prenat Diagn*. 2011;31:84-89.
7. Kagan KO, Sonek J. How to measure cervical length. *Ultrasound Obstet Gynecol*. 2015;45:358-362.
8. Retzke JD, Sonek JD, Lehmann J, Yazdi B, Kagan KO. Comparison of three methods of cervical measurement in the first trimester: single-line, two-line, and tracing. *Prenat Diagn*. 2013;33:262-268.
9. França C, Carraca T, Monteiro SB, et al. Inter- and intra-observer variability in cervical measurement by ultrasound in the first and second trimesters of pregnancy: does it matter? *J Perinat Med*. 2015;43:67-73.
10. Heath VC, Southall TR, Souka AP, Novakov A, Nicolaides KH. Cervical length at 23 weeks of gestation: relation to demographic characteristics and previous obstetric history. *Ultrasound Obstet Gynecol*. 1998;12:304-311.
11. Saltvedt S, Almstrom H, Kublickas M, Reilly M, Valentin L, Grunewald C. Ultrasound dating at 12–14 or 15–20 weeks of gestation? A prospective cross-validation of established dating formulae in a population of in-vitro fertilized pregnancies randomized to early or late dating scan. *Ultrasound Obstet Gynecol*. 2004;24:42-50.
12. Selbing A, Kjessler B. Conceptual dating by ultrasonic measurement of the fetal biparietal diameter in early pregnancy. *Acta Obstet Gynecol Scand*. 1985;64:593-597.
13. Kottner J, Audigé L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol*. 2011;64:96-106.
14. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8:135-160.
15. Bland JM, Altman DG. Measurement error proportional to the mean. *BMJ*. 1996;313:106.
16. Bland JM, Altman DG. Measurement error. *BMJ*. 1996;312:1654.
17. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86:420-428.
18. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound Obstet Gynecol*. 2008;31:466-475.
19. Kundel HL, Polansky M. Measurement of observer agreement. *Radiology*. 2003;228:303-308.
20. Cohen J. A coefficient of agreement for nominal scales. *Education Psychol Measure*. 1960;20:37-46.
21. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76:378-382.
22. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ*. 1992;304:1491-1494.
23. Kuhrt K, Smout E, Hezelgrave N, Seed PT, Carter J, Shennan AH. Development and validation of a tool incorporating cervical length

and quantitative fetal fibronectin to predict spontaneous preterm birth in asymptomatic high-risk women. *Ultrasound Obstet Gynecol*. 2016;47:104-109.

24. de Carvalho MH, Bittar RE, Brizot Mde L, Bicudo C, Zugaib M. Prediction of preterm delivery in the second trimester. *Obstet Gynecol*. 2005;105:532-536.
25. Esplin MS, Elovitz MA, Iams JD, et al. Predictive accuracy of serial transvaginal cervical lengths and quantitative vaginal fetal fibronectin levels for spontaneous preterm birth among nulliparous women. *JAMA*. 2017;317:1047-1056.
26. Grobman WA, Thom EA, Spong CY, et al. 17 Alpha-hydroxyprogesterone caproate to prevent prematurity in nulliparas with cervical length less than 30 mm. *Am J Obstet Gynecol*. 2012;207(390):e1-8.
27. Kuusela P, Jacobsson BO, Söderlund M, et al. Transvaginal sonographic evaluation of cervical length in the second trimester of asymptomatic singleton pregnancies, and the risk of preterm delivery. *Acta Obstet Gynecol Scand*. 2015;94:598-607.
28. van der Ven J, van Os MA, Kazemier BM, et al. The capacity of mid-pregnancy cervical length to predict preterm birth in low-risk women: a national cohort study. *Acta Obstet Gynecol Scand*. 2015;94:1223-1234.
29. Fonseca EB, Celik E, Parra M, Singh M, Nicolaides KH. Progesterone and the risk of preterm birth among women with a short cervix. *New Engl J Med*. 2007;357:462-469.
30. Hassan SS, Romero R, Vidyadhari D, et al. Vaginal progesterone reduces the rate of preterm birth in women with a sonographic short cervix: a multicenter, randomized, double-blind, placebo-controlled trial. *Ultrasound Obstet Gynecol*. 2011;38:18-31.
31. Norman JE, Marlow N, Messow C-M, et al. Vaginal progesterone prophylaxis for preterm birth (the OPPTIMUM study): a multicenter, randomised, double-blind trial. *Lancet*. 2016;387:2106-2116.
32. Winer N, Bretelle F, Senat M-V, et al. alpha-hydroxyprogesterone caproate does not prolong pregnancy or reduce the rate of preterm birth in women at high risk for preterm delivery and a short cervix: a randomized controlled trial. *Am J Obstet Gynecol*. 2015;212:485.e1-485.e10.
33. Souka AP, Pilalis A. Reproducibility of cervical length measurement throughout pregnancy. *J Matern Fetal Neonatal Med*. 2019;8:1-7.
34. Cruz-Melguizo S, San-Frutos L, Martínez-Payo C, et al. Cervical pessary compared with vaginal progesterone for preventing early preterm birth: a randomized controlled trial. *Obstet Gynecol*. 2018;132:907-915.
35. Goya M, Pratcorona L, Merced C, et al. Cervical pessary in pregnant women with a short cervix (PECEP): an open-label randomised controlled trial. *Lancet*. 2012;379:1800-1806.
36. Romero R, Conde-Agudelo A, Da Fonseca E, et al. Vaginal progesterone for preventing preterm birth and adverse perinatal outcomes in singleton gestations with a short cervix: a meta-analysis of individual patient data. *Am J Obstet Gynecol*. 2018;218:161-180.
37. Son M, Grobman WA, Ayala NK, Miller ES. A universal mid-trimester transvaginal cervical length screening program and its associated reduced preterm birth rate. *Am J Obstet Gynecol*. 2016;214(365):365.e1-365.e5.
38. Owen J, Yost N, Berghella V, et al. Mid-trimester endovaginal sonography in women at high risk for spontaneous preterm birth. *JAMA*. 2001;286:1340-1348.
39. Souka AP, Papastefanou I, Michalitsi V, et al. Cervical length changes from the first to second trimester of pregnancy, and prediction of preterm birth by first-trimester sonographic cervical measurement. *J Ultrasound Med*. 2011;30:997-1002.
40. Boelig RC, Feltovich H, Spitz JL, Toland G, Berghella V, Iams JD. Assessment of transvaginal ultrasound cervical length image quality. *Obstet Gynecol*. 2017;129:536-541.
41. Iams JD, Grobman WA, Lozitska A, et al. Adherence to criteria for transvaginal ultrasound imaging and measurement of cervical length. *Am J Obstet Gynecol*. 2013;209(365):365.e1-365.e5.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Kuusela P, Wennerholm U-B, Fadl H, et al. Second trimester cervical length measurements with transvaginal ultrasound: A prospective observational agreement and reliability study. *Acta Obstet Gynecol Scand*. 2020;99:1476–1485. https://doi.org/10.1111/aogs.13895