Final

# Statistical Analysis Plan

# Reproducibility of cervical length measurements in mid-pregnancy with vaginal ultrasound

2017-10-28

Approvals

| | |
|---|---|
| Name/Title: Nils-Gunnar Pehrsson / Senior Biostatistician | |
| ..... Signature: | ..... Date |
| Name/Title: Lil Valentin / Principal Investigator | |
| ..... Signature: | ..... Date |
| Name/Title: Ulla-Britt Wennerholm / Investigator | |
| ..... Signature: | ..... Date |

# Table of Contents

1 Study Details ..... 3
1.1 Introduction ..... 3
1.2 Study Objectives ..... 3
1.3 Design Study 1 "Live" ultrasound examination ..... 3
1.4 Design Study 2 "Video Clips" ..... 4
2 Study Variables ..... 4
2.1 Background Variables Study 1 Live ..... 4
2.2 Reproducibility Variables Study 1 Live (inter-rater agreement) ..... 4
2.3 Background Variables Study 2 Video Clips ..... 5
2.4 Reproducibility Variables Study 2 Video Clips (inter- and intra-rater agreement) ..... 5
3 Statistical Methodology ..... 6
3.1 Study 1 Live study ..... 7
**3.1.1 Inter-rater agreement and reliability. Continuous measurements** ..... 7
**3.1.2 Inter-rater agreement and reliability. Dichotomous measurements, Isthmus Yes/No** 8
**3.1.3 Measurement error. Intra-rater repeatability on the same occasion.** ..... 9
**3.1.4 Description of background variables Live study.** ..... 9
3.2 Study 2 Video Clips study ..... 9
**3.2.1 Inter-rater agreement and reliability. Continuous measurements** ..... 9
**3.2.2 Inter-rater reliability and agreement. Dichotomous measurements, Isthmus Yes/No** 10
**3.2.3 Intra-rater repeatability. Continuous measurements** ..... 10
**3.2.4 Intra-rater agreement and reliability. Dichotomous measurements, Isthmus Yes/No** 12
**3.2.5 Variance component analysis for the Clips study** ..... 12
**3.2.6 Description of background variables for the Clips study** ..... 12
4 Listing of Table, Figures and Listings ..... 13
4.1 Listing of Tables ..... 13
4.2 Listing of Figures ..... 13

# 1 STUDY DETAILS

## 1.1 Introduction

Measurement of cervical length in mid-pregnancy with ultrasound has been proposed as a screening method for preterm delivery. In a Swedish multicentre study with the title ”Prediktion av förtidsbörd: Screening av cervix med (vaginalt) ultraljud i andra trimestern hos asymptomatiska kvinnor med enkelbörd – en svensk multicenter observationsstudie” (“Prediction of preterm delivery: Screening with vaginal ultrasound in the second trimester in asymptomatic women with a singleton pregnancy- a Swedish multicentre observational study"), here referred to as the cervix study, the cervix will be measured with ultrasound in 11,000 pregnant women. The measurements are performed by specially trained midwives at 18+0 to 23+6 gestational weeks. It is important to know to what extent cervical length measurements are reproducible and reliable before they are introduced into clinical practice.

## 1.2 Study Objectives

To estimate the intra-rater and inter-rater repeatability/reproducibility, agreement and reliability of ultrasound measurements of cervical length at 18+0 to 23+6 gestational weeks.

## 1.3 Design Study 1 "Live" ultrasound examination

Patients who are eligible to be included in the cervix study are asked orally and through written information if they can accept to have their cervix examined with ultrasound not only by one midwife but by two midwives in the same scanning session. Thirty consecutive patients who consent are examined by one pair of examiners. The examinations are carried out as follows.

- Cervical length is measured with vaginal ultrasound in the same woman by two different examiners with the shortest possible interval (a few minutes) between the two examinations.
- The measurements are performed in exactly the same manner as in the cervix study, i.e. three measurements of each distance are taken following the instructions of the cervix study, and the presence or absence of isthmus is noted.
- The examiners take turn to measure first and second.
- The examiner's name must be visible on the saved ultrasound images to know who took the measurement.
- It is essential that the examiners do not know each other´s results. No frozen image - with or without measurements - may be left on the ultrasound screen.
- The investigators must not compare their images or written results during the course of the study (this may result in improved agreement and may bias results). When all data have been collected, the examiners are allowed to compare their images and results.
- All ultrasound images are saved digitally following the instructions for the "cervix study".
- One study pair, i.e. two examiners per participating center (Lund, Malmö, Solna, Huddinge, Gothenburg, Falun, Örebro) will be included in this Live study, Study 1. This means that the results of seven examiner pairs where both examiners work in the same center will be analysed.

### 1.4 Design Study 2 "Video Clips"

For the purpose of this study, 100 electronic video clips of consecutive ultrasound examinations of the uterine cervix in pregnant women are collected by one dedicated midwife measuring the cervix in the cervix study (as part of the study protocol of the cervix study, all ultrasound examinations of the cervix are documented with still images and often with video clips). These 100 clips are analysed by all the ultrasound midwives who perform cervical length measurements in the cervix study. The measurements are taken in the same way as in the cervix study, but only one measurement per distance is taken (not three repeated measurements per distance). The presence of isthmus is noted. Results are recorded in an excel file saved in the personal folder of each midwife on their hospital server. It can only be accessed by using the personal code of the midwife. A safety copy is saved on a personal USB stick of each midwife. Paper images of each measurement are printed and attached to a dedicated paper form for each clip.

To estimate intra-observer repeatability the same 100 video clips are re-analysed, but in a different order, by the same midwife at least 2 months later in the same manner as described above.

## 2 STUDY VARIABLES

### 2.1 Background Variables Study 1 Live

- Pregnant woman´s age
- Body mass index (from patient´s antenatal record)
- Parity
- Ethnicity
- Previous surgery on the cervix (cone biopsy of any type)
- Previous late miscarriage or preterm delivery
- Gestational age at the cervical length measurement included in the reproducibility study

### 2.2 Reproducibility Variables Study 1 Live (inter-rater agreement)

- A-B (length of the closed cervical canal) measured 3 times in the same session. Will be analysed for: a. All measurements, b. Measurements with Isthmus = no (i.e. B-C = 0) for both raters, c. Measurements with Isthmus = Yes (i.e. B-C >0) for both raters

- A-C (if no isthmus A-C= A-B) measured 3 times in the same session. Will be analysed for: a. All measurements, b. Measurements with Isthmus = Yes (i.e. B-C >0) for both raters

- Isthmus YES or isthmus NO noted once

- B-C measured 3 times in the same session (length of the isthmus zero not included, i.e. only if both midwives, i.e. both raters, have a measurement of B-C > 0 will the measurements be compared).Will only be analysed for: Measurements with Isthmus = Yes (B-C > 0 ) for both raters.

- (A-B)+(B-C) (closed cervical canal+isthmus) derived from A-B and B-C and measured 3 times in the same session. Length of the isthmus zero not included, i.e. only if both midwives, i.e. both raters, have a measurement of B-C >0, will the

measurements be compared. If no isthmus (i.e. B-C = 0) this distance is the same as A-C and the same as A-B. Will only be analysed for: Measurements with Isthmus = Yes for both raters (i.e. B-C >0 for both raters).

The distances to be measured are illustrated in Figure 1. Each distance is measured three times, and all three measurement results are recorded.

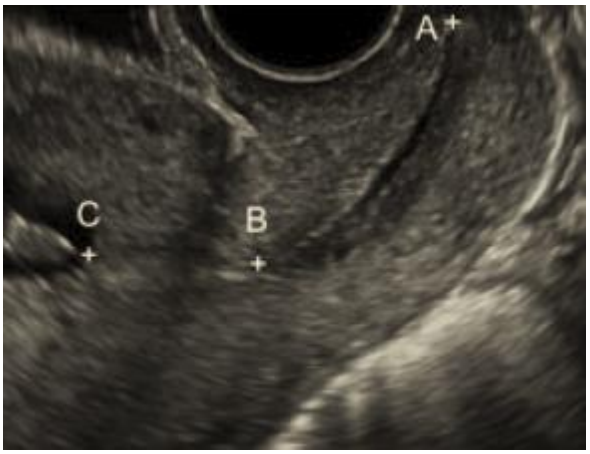


*Figure 1. Ultrasound image of the cervix obtained with vaginal ultrasound. A, outer cervical os; B, inner cervical os; C, the virtual inner cervical os created by the opposition of the anterior and posterior isthmus. The distances measured are: A-B (endocervical length), B-C (isthmus), A-C ("as the crow flies" or "bee line"). In addition [(A-B)+ B-C)] is calculated*

### 2.3 Background Variables Study 2 Video Clips

- Pregnant woman´s age
- Body mass index (from patient´s antenatal record)
- Parity
- Ethnicity
- Previous surgery on the cervix (cone biopsy of any type)
- Previous late miscarriage or preterm delivery
- Gestational age at the cervical length measurement included in the reproducibility study

### 2.4 Reproducibility Variables Study 2 Video Clips (inter- and intra-rater agreement)

- A-B (length of the closed cervical canal) measured once at two different assessment sessions at least 2 months apart.
  *Will be analysed for: a. All measurements b. Measurements with Isthmus = no (B-C = 0) c. Measurements with Isthmus = Yes (B-C>0)*.

- A-C (if no isthmus A-C= A-B) measured once at two different assessment sessions at least 2 months apart
  *Will be analysed for: a. All measurements, b. Measurements with Isthmus = Yes (B-C = >0)*

- Isthmus YES or isthmus NO noted once at two different assessment sessions at least 2 months apart

- B-C measured once at two different assessment sessions at least 2 months apart (B-C zero not included; for intra-rater agreement: only if the same rater has a measurement of B-C >0 at both rating sessions will the measurements of B-C be compared between the repeated measurements by the same observer)
  *Will only be analysed for: Measurements with Isthmus = Yes (B-C>0)*

- (A-B)+(B-C) (closed cervical canal+isthmus) derived from A-B+B-C and measured once at two different assessment sessions at least 2 months apart. (B-C zero not included; for intra-rater agreement, only if the same rater has a measurement of B-C >0 at both rating sessions will this measurement be compared between the repeated measurements by the same observer).
  *Will only be analysed for: Measurements with Isthmus = Yes (B-C>0)*

## 3 STATISTICAL METHODOLOGY

One patient can participate only once in the reproducibility study. If a patient has been examined twice in the reproducibility study (by mistake) the measurement taken on the first examination occasion is used in our statistical calculations. If there are missing data for the first examination but complete data for the second, then the second examination will be used. If there are missing data for both examination occasions, the examination with the most complete data will be used.

PLEASE OBSERVE All cervical lengths will be rounded to nearest integer (mm) if decimal integer has been entered into the database. This will be made prior to any other calculations.

Continuous data will be presented with mean, standard deviation, median, min, max and number of subjects. Categorical data will be presented with number of subjects and percent.

For all pair-wise inter-rater comparisons in the live study and all intra-rater comparisons in the Clips study Bland-Altman plots will presented (that is a plot of differences between measurements against the mean of the measurements). **Figure 1.1.1-1.1.x and 2.2.1 -2.2.x**

Before analysing the results of the LIVE and Clips studies we want to assess the relationship between intra- and inter-rater differences and the magnitude of the measurement values.

*Inter-rater:* For Live reproducibility we will do this by plotting the absolute differences against the mean measurement result of the two raters in the same pair in one and the same plot for all seven pairs, with each pair being represented with a different color. In addition we will make one plot for each pair. Spearman correlation coefficient will be calculated for each pair of raters and for all raters together. If there is an obvious tendency of increasing absolute differences with increasing mean of the measurements we will make a logarithmic transformation of the data. We will then plot log differences and against log mean to verify that the correlation disappears. We will make these analyses for A-B (mean of three measurements) and B-C (mean of three measurements). The results for A-B will be applied to all measures of A-B, A-C and (A-B)+(B-C) for both Live study and Clips study, and the results for B-C will be applied to all measurements of B-C for both studies. If the absolute differences increase with the measurement values, log transformation and re-transformation will be used to present the differences as a ratio between the two examiners´ measurement results both for the LIVE study and the clips study (**Figure 1.3.1-1.3.x and Figure 1.4**).

*Intra-rater:* For the Clips study we plot intra-individual (II) SD (SD = absolute difference / sqrt (2)) against the mean of the rater´s two measurements both for each rater and for all 16 raters together. Spearman correlation coefficient will be calculated for each rater and for all raters together. If there is an obvious tendency of increasing IISD with increasing mean of the measurements we will make a logarithmic transformation of the data. We then plot log differences against log mean to verify that the correlation disappears. We make these analyses for A-B and B-C. The results for A-B will be applied to measurements of A-B, A-C and (A-B)+(B-C) in the Clips study and the result for B-C will be applied to all B-C measurements in the Clips study **(Figure 2.3.1-2.3.x).**

Depending on the results of the above analyses we will decide if to show results for continuous variables as differences in mm or as a ratio.

## 3.1 Study 1 Live study

In study 1 "Live study" the inter-rater agreement and reliability between two examiners within each centre is studied.

### *3.1.1 Inter-rater agreement and reliability. Continuous measurements*

For variables see Section 2.2.

For each of these measurements we make the inter-rater agreement and reliability analyses on:

- The minimum value of the three measurements
- The maximum value of the three measurements
- The mean value of the three measurements

**AGREEMENT (**how much do measurements differ between raters in mm or in percent (ratio))

For each of the seven rater pairs the measurements of one rater are plotted against those of the other rater (at least for the most important variables). (**Figure 1.8.1 -1.8.x**)

For each study pair, a Bland-Altman plot is made to describe how the differences in measurement results are related to the measurement values (**Figure 1.1.1-1.1.x**).

For each study pair, the mean and SD, median, minimum and maxium of the measurement values are calculated. The distribution of the differences between the two raters is presented as Mean, SD, Median, Minimum and maximum **(Table 1.2.1 – 1.2.7).**

When calculating the differences, the values from the rater with the lowest mean (of three) measurement values for A-B (named rater 2) are subtracted from those of the rater with the highest mean (of three) measurement values for A-B (named rater 1).

In agreement with what is stated under 3 above, if the absolute differences do not increase with the mean of the measurement values, then the limits of agreement (mean difference +/- 1.96 SD; mean difference minus 1.96 SD is the lower limit of agreement; mean difference +1.96SD is the upper limit of agreement) for the individual differences is the main result from this inter-rater agreement study. In other words, if we have a measurement from one of the examiners, then from the limits of agreement we can estimate the interval for the other examiner's measurement for 95% of future observations. The 95% CI for the mean difference will also be calculated. From this we can estimate if there are any systematic differences between the two examiners (if the 95%CI does not include zero there is a systematic difference).

In agreement with what is stated under 3 above, if the absolute differences increase with the mean of the measurement values then we take the logarithm for all values and calculate limits of agreement for those values. The difference and the limits of agreement are then antilogarithmed to get the limits of agreement expressed as a ratios. If the mean ratio is 1.20, it means that the results of one examiner are 20% higher than those of the other. A 95% CI for the mean ratio of the geometric means will also be calculated, from which we can estimate if there is any systematic difference between the two examiners (if the 95% CI does not include 1, there is a systematic difference).

A forest plot for mean difference (or mean ratio) with limits of agreement and mean difference (or mean ratio) with 95%CI will be presented for all sites for each analysed variable separately. In the forest plots all mean differences will be turned positive (or if to present results as ratios all ratios will be shown as $\geq$1) and the limits of agreement adjusted accordingly (**Figure 1.5.1-1.5.x**).

A Bland-Altman plot with difference on the Y-axis and mean measurement values on the x axis for all observations from all sites with different colours for each centre will be presented, if informative (in this plot all negative differences will be shown as positive differences) (**Figure 1.2**).

**RELIABILITY**
The Intraclass correlation coefficient (ICC) with 95% CI will be calculated as a measure of reliability. For calculation of ICC, ANOVA (analysis of variance) is used. Both a two-way mixed model (absolute agreement) and two-way random model (absolute agreement) will be used: we want to check the reliability of our results for the raters in our study (mixed model) but we also want to generalize to other raters (random model).

**Tables**
One table for each of the seven rater pairs (**Table 1.2.1 to 1.2.7**) will be created showing the mean, SD, median, minimum and maximum measurement values for rater 1 and rater 2 (for definition of rater 1 and rater 2, see above) and for the differences between the two raters for all variables analyzed. The 95%CI for the mean difference, the lower and upper limits of agreement, and ICC for all variables analyzed will also be shown. If differences between raters are presented as ratios instead of mm, then a column showing mean, SD, median, minimum and maximum ratio (value of rater 2 divided by value of rater 1) will be added, and the 95%CI for the mean ratio as well as limits of agreement for the ratios (instead of for difference in mm) will be shown. In an additional Table the distribution of the measurement results will be shown summarizing the results of all raters (**Table 1.3**).

#### *3.1.2 Inter-rater agreement and reliability. Dichotomous measurements, Isthmus Yes/No*

The inter-rater analysis of isthmus consists of calculation for each pair of examiners: percent agreement, percent positive agreement, percent negative agreement (see Kundel and Polanski) and Cohen's kappa (reliability) with 95% CI. These statistics will be presented separately for all seven pairs of examiners but also summarized as mean, median, minimum and maximum of: percentage agreement, percentage positive agreement, percentage negative agreement, and Cohen´s Kappa over all seven pairs **(Table 1.4).**

The results for interrater agreement and reliability with regard to presence or absence of isthmus will also be illustrated in **Figures** showing on the "y-axis" percent agreement, positive agreement and negative agreement, and Cohen´s Kappa for each of the seven rater pairs ("x-axis") in the same Figure **(Figure 1.6.1 – 1.6.x).**

#### *3.1.3 Measurement error. Intra-rater repeatability on the same occasion.*

Calculation of intra-individual (within subject) SD and INTRA-rater coefficient of variation for each rater for all continuous variables in section 2.2. A plot of intra-rater intra-individual (within-subject) SD against mean of the three measurements will be created for each rater. (Figure 1.7.1 -1.7.x). The intra-rater intra-individual (within-subject) SD and intra-rater coefficient of variation will be presented for each rater and the distribution of individual (within-subject*) SD and intra-rater coefficient of variation presented over all raters (Tabell 1.5).*

#### *3.1.4 Description of background variables Live study.*

The background variables presented in section 2.1 will be described by center (Table 1.1).

### 3.2 Study 2 Video Clips study

In study 2, the Video Clips study, the inter-rater agreement and reliability between all 16 raters from all centres will be studied and the intra-rater repeatability and reliability of all raters will be studied.

#### *3.2.1 Inter-rater agreement and reliability. Continuous measurements*

The variables to be analyzed are presented in section 2.2. However, there is only ONE measurement per distance, so mean, minimum and maximum value for the same distance does not apply here. *For the analysis of inter-rater agreement and reliability, the results of the <u>first analysis round of the clips</u> will be used.*

Inter-rater AGREEMENT
Systematic differences between raters are analysed with a two-way-ANOVA with video clips as block effects (**Table 2.2**)

*Acording to Jones et al.(2011)), a* plot is constructed with each rater's difference from the mean of the measurement results of all raters on the Y-axis and the mean of all raters´ measurements for the individual patient (video clip) on the x- axis. The limits of agreement with the mean are presented in the plot. The plot should have different symbols or colours for each rater. The limits of agreement with the mean show by how much an individual rater´s measurement can differ from the mean of all raters´ measurements (**Figure 2.1**).

The distribution of mean differences of the 120 pairs, upper limits of agreement and lower limits of agreement will be described as the mean, SD, median, minimum and maximum. Doing so, all mean differences will be shown as positive differences and limits of agreement will be adjusted correspondingly. If results are shown as ratios, all ratios will be shown as ratios $\geq$ 1.0 and limits of agreement will be adjusted accordingly. (**Table 2.3;** this Table will also show the measurement values, the distribution of the measurement values being calculated using all values, **see Table 2.3**).

Inter-rater RELIABILITY
IntraClass Correlation Coefficient (ICC) will be calculated from the ANOVA as a measure of reliability. We will use both a mixed model (absolute agreement) and a

random model (absolute agreement), the former to estimate reliability in our own study, the latter to generalize our results to other raters (results will be shown in **Table 2.3**).

#### *3.2.2 Inter-rater reliability and agreement. Dichotomous measurements, Isthmus Yes/No*

AGREEMENT
The distribution of percentage agreement, percentage positive agreement, and percentage negative agreement over all 120 pairs of raters will be presented as Mean, SD, Median, minimum and maximum (**Table 2.4**).
RELIABILITY
Reliability regarding the occurrence of isthmus is expressed as Fleiss Kappa with 95% CI, CI calculated by jack-knife technique. The distribution of Cohens Kappa over all 120 pairs of raters will be presented as Mean, SD, Median, minimum and maximum (**Table 2.4**).

#### *3.2.3 Intra-rater repeatability. Continuous measurements*

For variables see Section 2.2. However, there is only ONE measurement per distance, so mean, minimum and maximum value for the same distance does not apply here.

For each of the 16 raters the measurements at the first assessment round are plotted against those of the second assessment round (**Figure 2.4.1-2.4.x**)

AGREEMENT

For each rater a Bland-Altman plot is made to describe how the differences in measurement results between the first and second measurement are related to the measurement values (**Figure 2.2.1 -2.2-x**). In addition SD is plotted against mean measurement values (**Figure 2.3.1-2.3.x**) for A-B and B-C, see under 3 (statistical methodology). If SD increases with the measurement values log transformation and re-transformation are used to instead present the differences as a ratio between the raters' measured values (second measurement divided by first measurement).

For each rater, the distribution of the first and second measurement values is described as mean, SD, median, minimum and maximum, and the distribution of the differences between the two measurements is presented as Mean, SD, Median, Minimum and maximum (**Table 2.5.1-2.5.16**).

If the absolute differences do not increase with the mean of the values, then the within individual (Intra-Individual) SD (IISD) for each rater is the main result from this intra-rater repeatability study. The difference between a subject's measurement and the true value would be expected to be less than 1.96 * IISD for 95% of observations. The repeatability is $\sqrt{2}$ * 1.96 * IISD, i.e. the difference between two measurements for the same subject is expected to be less than the repeatability for 95% of future pairs of observations. The 95% CI for the mean difference will also be calculated. From this we can see if there are any systematic differences between the first and second measurements. (**Table 2.5.1-2.5.16**).

If the intra individual SD increases with the mean of the values then the main result is the intra-rater *coefficient of variation*. Let $s_w$ be intra-individual SD in the log scale. Let $a_{sw}$ be the antilog of $s_w$. The true value for 95% of the observations should fall between the measured value divided by $a_{sw}$ ** 1.96 and the measured value multiplied with $a_{sw}$ ** 1.96. (**Table 2.5.1-2.5.16**).

RELIABILITY

The Intraclass correlation coefficient (ICC) with 95% CI will also be calculated. For calculation of ICC, ANOVA (analysis of variance) is used, two-way mixed model, absolute agreement (as an estimate of reliability in our study) and two-way random model, absolute agreement (as an estimate of reliability in general). (**Table 2.5.1-2.5.16**).

In a Summary Table (**Table 2.6**) The distributions of intra-rater mean difference (or ratio), within subject SD's (CV's), repeatability and ICC will be presented over all raters as mean, SD, median, minimum, maximum.

### *3.2.4 Intra-rater agreement and reliability. Dichotomous measurements, Isthmus Yes/No*

AGREEMENT
The intra-rater analyses of isthmus (yes or no) consist of calculation of percent agreement, percent positive agreement, percent negative agreement. The distribution of percent agreement, percent positive agreement, percent negative agreement over all 15 raters will be presented as mean, SD, median, minimum and maximum (**Table 2.7)**

RELIABILITY
Cohen's kappa with 95% CI for each rater. The distribution of Cohen's kappa over all 15 raters will be presented as mean, SD, median, minimum and maximum (**Table 2.7**)

The intra-rater agreement and reliability results will also be shown in a summary Figure (**Figure 2.5.1-2.5.x**): on the "y-axis" percent agreement, positive agreement, negative agreement, and Cohen´s Kappa for each of the 16 raters (rater on "x-axis")

### *3.2.5 Variance component analysis for the Clips study*

Model:

The response $Y_{ijkl}$ for subject i by rater j at time k and measurement l is given by

$$Y_{ijkl} = \mu + X_i + O_j + H_{ij} + W_{ijk} + {}_{ijkl},$$

where $X_i$ i is the subject effect, $O_j$ the rater effect, $H_{ij}$ the subject*rater effect (heterogeneity), $W_{ijk}$+ $\varepsilon_{ijkl}$ the subject*rater*time effect (heterogeneity within subject) and $\varepsilon_{ijkl}$ is the measurement error of by a single rater on a single subject at a specific time point. $E_{ijkl}$ cannot be estimated separately because there is only one measurement at each time point. An ANOVA table will be constructed where Source of variation, Degrees of freedom and Expected mean sum of squares will be presented. From the output of the ANOVA model all the above variance components can be estimated. (**Table 2.8**)

### *3.2.6 Description of background variables for the Clips study*

The background variables presented in section 2.3 will be described for the women providing the clips (n = 93 after removal of 7 examinations from the same woman) (**Table 2.1**)
Continuous variables will be described by mean, SD, median, minimum and maximum and categorical variables with number and percentages.

# 4 LISTING OF TABLE, FIGURES AND LISTINGS

## 4.1 Listing of Tables

| Table Number | Table Title |
|---|---|
| LIVE study | |
| 1.1 | Live Study. Background variables for the 7 inter-rater reliability study populations with center as columns |
| 1.2.1-1.2.7 | Live Study. Inter-Rater agreement and reliability for continuous variables, one Table for each center (rater pair) |
| 1.3 | Summary table for inter-rater agreement and reliability (ICC) for the 7 centers (rater pairs) |
| 1.4. | Live Study. Inter-rater agreement and reliability for presence/absence isthmus for each center (rater pair) including a summary for the 7 centers |
| 1.5 | Live study. Distribution of INTRA-rater SD and INTRA-rater coefficient of variation |
| CLIPS study | |
| 2.1 | Video Clips Study. Background variables for the 93 women included in the Video Clips Study |
| 2.2. | Video Clips Study: Comparison of the means of the 16 raters (ANOVA) |
| 2.3 | Video Clips Study: Summary Table for Inter-Rater agreement: limits of agreement with the mean, distribution of mean difference/ratio and lower and upper limit of agreement, and ICC |
| 2.4 | Video Clips Study: Inter-Rater agreement, specific agreement and reliability for presence/absence isthmus. Distribution of agreement, specific agreement and Cohens kappa. Fleiss Kappa |
| 2.5.1-2.5.16 | Video Clips Study: **Intra**-Rater agreement and reliability for continuous variables, one Table for each of 16 raters |
| 2.6 | Video Clips Study: **Intra**-Rater agreement for continuous variables: summary for all 16 raters |
| 2.7 | Video Clips Study: **Intra**-Rater agreement and reliability with regard to presence/absence isthmus for all raters and with summary |
| 2.8 | Video Clips Study: Results Variance Components Model |

## 4.2 Listing of Figures

| Figure Number | Figure title |
|---|---|
| 1.1.1-1.1.x | Live study Inter-rater Bland-Altman plots Difference vs mean for each measurement for each pair of raters |
| 1.2 | Live study Inter-rater Bland-Altman plots Difference vs mean for each measurement for all pair of raters in one single plot |
| 1.3.1-1.3.x | Live study Inter-rater Bland-Altman plots Absolute Difference vs mean, A-B and B-C, for each measurement for ***each*** pair of raters |
| 1.4 | Live study Inter-rater Bland-Altman plots Absolute Difference vs mean, A-B and B-C, for each measurement for ***all*** pairs of raters |
| 1.5.1-1.5.x | Live study Forest plots for Limits of agreement and mean diff (or ratio) with 95%CI |
| 1.6.1-1.6.x | Live study. Presence or absence of isthmus: on the “y-axis” percent agreement, positive agreement, negative agreement, and Cohen´s Kappa for each of the seven rater pairs (rater pair on “x-axis”) in the same Figure |

| | |
|---|---|
| 1.7.1-1.7.x | Live study. Plot of INTRA-rater SD against mean of (three) repeated measurements |
| 1.8.1.1.8.x | For each of the seven rater pairs the measurements of one rater are plotted against those of the other rater |
| CLIPS study | |
| 2.1 | Video Clips Plots according to Jones et al |
| 2.2.1-2.2.x | Video Clips Intra-rater Bland-Altman plots Difference vs mean |
| 2.3.1-2.3.x | Video Clips Intra-rater Plot Intra individual SD vs mean |
| 2.4.1-2.4.x | Video Clips Plot of measurements at the first assessment round against those of the second for each of the 16 raters |
| 2.5.1-2-5-x | Video Clips. Presence or absence of isthmus: on the “y-axis” percent agreement, positive agreement, negative agreement, and Cohen´s Kappa for each of the 16raters (rater on “x-axis”) in the same Figure |