

Title

Individual ovarian volumes obtained from 2D and 3D ultrasound lack precision

Running Title

Precision of ovarian volume measurements

Authors

Brett S (1), Bee N (2), Wallace WHB (3), Rajkhowa M (1), Kelsey TW (4) *.

Addresses

1 Assisted Conception Unit, Ninewells Hospital and Medical School, Dundee,

DD1 9SY

2 College of Medicine & Veterinary Medicine, The University of Edinburgh, EH16 4TJ

3 Royal Hospital for Sick Children, 17 Millerfield Place, Edinburgh, EH9 1LF.

4 School of Computer Science, University of St. Andrews, KY16 9SS

*To whom all correspondence should be addressed at:

Tom Kelsey, School of Computer Science, University of St.

Andrews, KY16 9SS

E-mail: tom@cs.st-and.ac.uk

Abstract

Objective – To assess the precision of individual ovarian volume measurements by 2D and 3D transvaginal ultrasound scan.

Study Design – Transvaginal 2D and 3D ultrasound examinations were performed on 49 women attending a tertiary centre for investigation or treatment for subfertility. Two observers calculated ovarian volume, using both the prolate ellipsoid formula and virtual organ computer-aided analysis (VOCAL), with rotation steps of 30 degrees (3D-30).

Results – For the four comparisons (inter- and intra-observer; 2D and 3D-30) we obtained intraclass coefficients of 0.97 to 0.98; and standard errors ranging from 17% to 14% (for inter-observer 2D and intra-observer 3D, respectively). The corresponding Coefficients of Repeatability ranged from 33% to 28%.

Conclusions - Measurement of transvaginal ovarian volumes using both 2D and 3D ultrasound is imprecise for individuals. The imprecision is greater for lower ovarian volumes, which may be important in clinical practice. The average of two or more measurements is likely to be more accurate than a single measurement.

Key words

Ovarian volume/IVF/2D/ 3D Ultrasound

Introduction

A screening tool which could effectively predict ovarian reserve would have the potential to assist women in making informed decisions about their fertility. A clinical application could be as an aid for prediction of ovarian response to stimulation with exogenous gonadotrophins in in-vitro fertilisation treatment (IVF). A highly significant correlation between primordial follicle population and ovarian volume has been reported by several authors (Lass, 1997a; Wallace, 2004) and it has been proposed that transvaginal estimation of ovarian volume may be used to determine ovarian reserve for an individual woman (Wallace, 2004).

In order for a screening tool based on ovarian volumes to be clinically useful, knowledge of the precision of the measurement technique used is essential. In this study we estimate the precision of ovarian ultrasound measurement by calculating Intraclass correlation coefficients (ICC), standard errors and coefficients of repeatability. Analysis of these three statistics allows us to estimate the precision of 2D and 3D ultrasound measurements not only for populations of subjects but also for individual measurements, which is perhaps of more clinical relevance.

Coefficients of repeatability are twice the standard error, as defined by the British Standards Institution (1979), so that two repeated measurements are expected to be within this percentage for 95% of subjects (Bland, 1986; Bland, 1999). An ideal screening technique would have ICC's close to 1 and standard errors close to 0.

Intraclass correlation coefficients (ICCs) are routinely reported in studies involving clinical measurements (Bland, 2000), and have previously been used in studies related to ovarian volume (Higgins, 1990; Kyei- Mensah, 1996a; Jarvela, 2003; Merce, 2005) Previous studies involving ICC estimation have shown that both 2D and 3D transvaginal ultrasound measurements of ovarian volume are highly reproducible. Published inter- and intraobserver ICCs for 3D measurements range from 0.95 -1.0 (Kyei-Mensah, 1996a; Jarvela 2003; Merce, 2005) with Higgins reporting an interobserver ICC of 0.96 for 2D measurements (Higgins, 1990). Values for ICCs above 0.75 are said to be acceptable (Burdock, 1963). Therefore, both 2D and 3D ultrasound measurements provide valid estimates of the true volume of the ovary

It has been suggested that 3D estimations of ovarian volume are superior to 2D (Bonilla-Musoles, 1995; Kyei-Mensah, 1996b; Raine-Fenning, 2003a), although this may depend on the technique used. Few studies exist which compare estimated volumes on scanning with actual measurements (Saxton, 1990; Bonilla- Musoles, 1995; Kyei- Mensah, 1996b; Raine-Fenning, 2003b).

The aim of this study is to determine the precision of ovarian volume measurements by 2D and 3D transvaginal ultrasound scan using three appropriate statistics: ICC, standard error, and coefficient of repeatability and therefore quantify the precision of measurements of ovarian volume for individual women rather than for populations.

Materials and Methods

Ethical approval was obtained from the local ethics committee. Forty-nine subjects were recruited from women attending for investigation of sub-fertility or undergoing assisted conception treatment at a single tertiary centre between January and May 2006. Informed consent was obtained from each subject, prior to their enrolment in this study. The patients were scanned at random stages of the cycle.

All ultrasound scans were performed by a single operator using a Voluson 730-Pro machine and a 7.5 MHz transvaginal probe. One subject had previously undergone a unilateral oophorectomy. We excluded 7 further ovaries due to inadequate visualisation. For each remaining ovary (n = 90), a set of 2-D images was stored to allow measurement of the standard three planes used for volume calculation. A 3-D data set was then acquired using a predefined probe program designed to optimise image quality. Data generated was stored for later analysis using 4D View (GE Healthcare).

Analysis was performed by two observers. An anonymised copy of the original data was given to the second observer to allow simultaneous independent data analysis. The standard maximal three diameters to calculate ovarian volume were obtained from captured 2-D images. 2D ovarian volume was calculated using the prolate ellipsoid formula ($V = D1 \times D2 \times D3 \times 0.523$); 3-D volume calculations were performed independently, with no access to 2D results.

Virtual organ computer-aided analysis (VOCAL) was used to conduct 3-D rotational measurements of ovarian volume using rotation steps of 30° in both the A- plane (longitudinal) and C-

plane (coronal). The longitudinal images (A-plane) represents the original data with all other images (transverse and coronal) obtained using reconstructed data. In this study the results obtained using the A-plane are presented.

Each dataset was independently measured on two separate occasions two months apart by each observer without recourse to initial volume calculations. The sample size requirements for reliability and precision studies are given as equation 12 and table II in (Walter, 1998). In order to achieve adequate power for this study, a minimum of 58 ovaries was required.

Six datasets (1st 2D observer 1, 2nd 2D observer 1, 2D observer 2, 1st 3D-30 observer 1, 2nd 3D-30 observer 1 and 3D-30 observer 2) were obtained, and four comparisons made: intra- and inter observer for both techniques. For each within-subject comparison, the differences, log-adjusted differences, and ratios were obtained. The standard error was taken to be the standard deviation of the ratios – as recommended in (Bland, 1999) - expressed as a percentage. For standard errors to be a suitable statistic, the change in mean between pairs of measurements should be small (i.e. within a few percent), so this value was also calculated. The ICC used was the ratio of the standard error to the mean between-subject standard deviation of the two measurements (derived by weighting the variances by their degrees of freedom) (Bland, 2000).

Results

The 49 subjects had a mean age of 33 years (SD 5.4 years, minimum 21 years, maximum 43 years). 27 subjects were not on

any form of hormonal treatment at the time of the study; the remainder were at varying stages of assisted conception treatment.

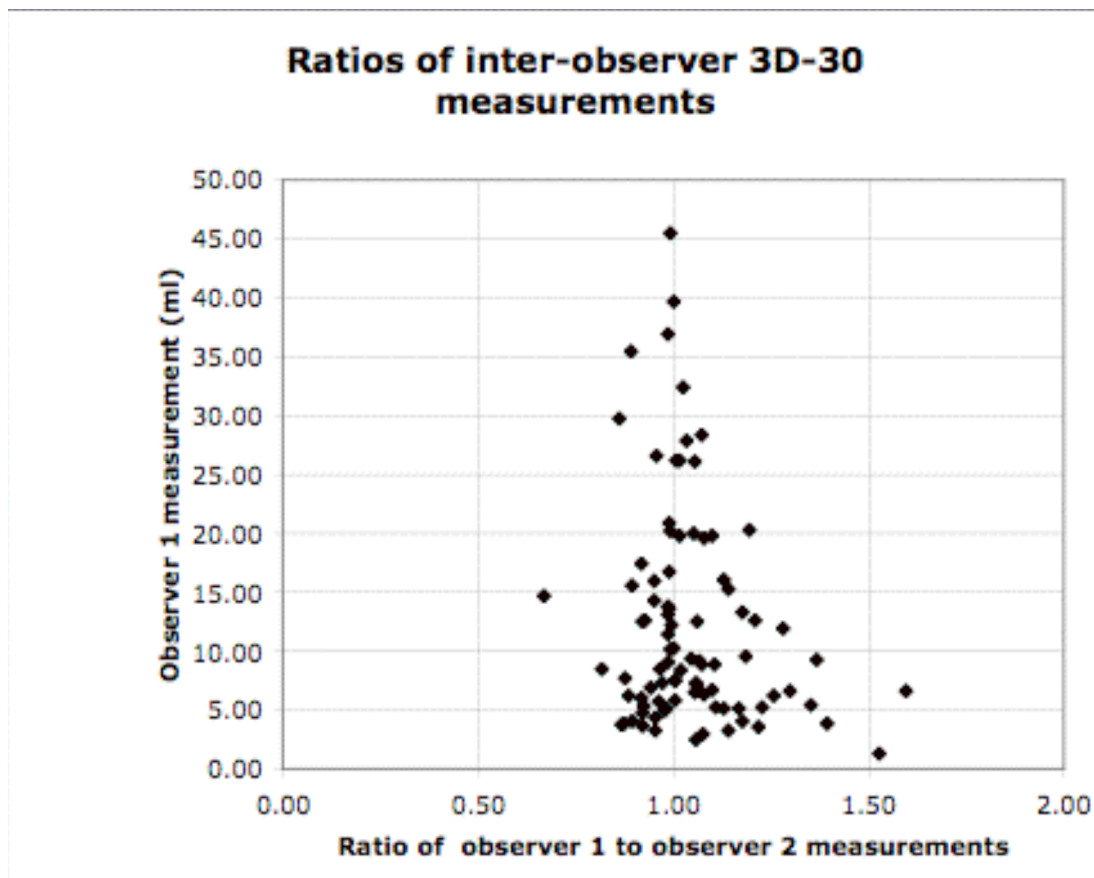
We performed four comparisons. For both 2D and 3D 30° VOCAL measurement techniques we compared repeated measures by the same observer (intraobserver repeatability) and measures by two different observers (interobserver reliability). Since, in all four cases, the differences had log-normal distributions, statistical analysis was performed using log-transformed data, with results presented after antilogarithmic transformation. Our results are set out in the Table.

All three precision-related statistics were better for 3D-30 estimates, but this improvement did not reach statistical significance. Relative errors increase with reduced ovarian volume, as shown in the Figure. Analysis restricted to ovaries with mean estimated volumes less than 15 ml led to slightly lower ICCs, with no corresponding improvement in the standard errors.

Table 1 – Summary of precision results.

| Statistic | Intra-observer | | Inter-observer | |
|----------------------------------|----------------|--------|----------------|--------|
| | 2D | 3D-30° | 2D | 3D-30° |
| Change in mean (%) | <3% | <3% | <1% | <3% |
| Standard error (%) | 14.9% | 14.2% | 16.7% | 14.6% |
| Coefficient of Repeatability (%) | 39.8% | 28.4% | 33.4% | 29.2% |
| Intraclass r | 0.98 | 0.99 | 0.98 | 0.99 |
| Lower 95% conf. limit | 0.97 | 0.99 | 0.96 | 0.98 |
| Upper 95% conf. limit | 0.99 | 0.99 | 0.99 | 0.99 |

Figure 1



Discussion

This is the first study to compare the precision of individual estimates of ovarian volume using both 2D and 3D ultrasound techniques. Our results show that both techniques are repeatable for populations, and our ICCs are similar to those obtained in other studies (Higgins, 1990; Kyei-Mensah, 1996a; Jarvela, 2003; Merce, 2005). However, it is important to remember that ICCs depend on the variance within the study population and that account must also be taken of the absolute variance in order to determine if a measurement method yields consistent results. The precision of a technique is important for individual measurements,

as an imprecise technique means that the true value lies within a large range and therefore hinders its predictive value. It is possible for any measurement technique to have a high estimated ICC, but relatively large standard errors. This will occur when relatively large variances from overestimation balance out similar variances in underestimation. In this case, the technique will be reproducible for populations, since the average of two repeated measurements is expected to be close to the true value. Since a high ICC can mask a high standard error, we derive and analyse both statistics. Bland-Altman coefficients of repeatability are closely related to the standard error of the technique, and allow us to estimate how close two measurements taken from the same scan are likely to be.

We report standard errors ranging from 14% to 17%, suggesting that individual measurements are imprecise. A repeated estimate from the same data is expected to be within 28% of the first estimate for 95% of subjects, when the standard error is 14%. In clinical terms this suggests that a repeat estimate of an ovary originally calculated at 4 mls would be within the range 2.2 - 5.1 mls.

Our results are in broad agreement with studies that assessed ultrasound measurement of the cervix in pregnant women (Rovas, 2005; Valentin, 2002). These studies reported good or excellent inter- and intraclass correlation, despite relatively high standard errors. Valentin & Bergelin (2002) reported considerable differences (7-12mms) in measurements of cervical length in a few women. Rovas et al. (2005) concluded that only large changes in cervical volume would be likely to be detected using current technology.

There are several examples of controversies and conflicting evidence based on ultrasound ovarian measurement. For example, Syrop reported that ovarian volume correlates with the number of retrieved oocytes after controlled ovarian stimulation (Syrop, 1999), but this result was not confirmed in a similar study (Tomas,1997). A further example is the finding that ovarian volume changes significantly throughout the cycle reported by Christensen (Christensen, 1997), whereas no intra-cycle variation in volume was detected in a later study (Oppermann, 2003). Our results suggest that such inconsistencies and discrepancies in the literature may be explained by the lack of precision in ultrasound measurement.

Research using 2D ultrasound estimations of either mean ovarian volume (Lass, 1997b; Sharara, 1999) or smallest ovary (Syrop, 1999) suggest a higher cancellation rate in patients whose estimated ovarian volume is < 3mls. Schild et al (2001) reported that 9.9% of patients in their study had a minimum unilateral ovarian volume of 3mls or less. Our results would suggest that for an individual woman the imprecision of 2D and 3D ultrasound in estimating ovarian volume lowers the likelihood of finding a single clinically useful cut-off value for ovarian volume below which treatment could confidently be withheld. Indeed, in their prospective cohort analysis, Frattarelli et al (2004) suggested that there was no absolute mean ovarian volume which could accurately predict cycle cancellation. In a systematic review of current literature, Broekmans et al (2006) concluded that measurement of ovarian volume showed a modest predictive accuracy for predicting poor response to stimulation but no clear accuracy in the prediction of treatment outcome. The ability to

determine in advance women who will have a poor response to ovarian stimulation, could aid in individualising management and lower the significant emotional and financial cost of cancelled cycles.

The use of 3D ultrasound techniques does not appear to significantly increase the predictive power of ovarian volume as a screening tool for response to ovarian stimulation. Jayaprakasan et al (2007) reported that measurement of mean ovarian volume using either '2D equivalent' or 3D ultrasound techniques was a poor predictor of the number of oocytes retrieved. In this study, as with ours, there was a trend towards increased precision with 3D volume calculations, but this did not reach statistical significance.

This study adds to our present understanding of the role of ovarian volume measurement in reproductive medicine. We have confirmed that the measurement of ovarian volumes by both 2D and 3D-30° techniques leads to high inter- and intra class correlation. However, both methods suffer from a lack of precision, which is greater for lower ovarian volumes. This has important implications for clinical practice, and may explain inconsistencies in the literature for studies in which the measurement of ovarian volume is an important factor. The results of this study suggest that individual measurements of ovarian volume are imprecise and therefore of doubtful clinical value. The possibility of measurement error should be taken into account when making clinical decisions and an average of multiple measurements (either by the same observer, or by different observers) is likely to give a more accurate estimate of the true ovarian volume.

References

Bland JM, Altman DG. 1986 Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. i, 307-310.

Bland JM, Altman DG. 1999 Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 8, 135-160.

Bland M. 2000 *An Introduction to Medical Statistics*. 3rd Ed. Oxford,

Bonilla-Musoles F, Raga F, Osborne NG. 1995 Three-Dimensional ultrasound evaluation of ovarian masses. *Gynecologic Oncology*, 59,129-135.

Burdock E I, Fleiss JL, Hardesty AS. 1963 A new view of interobserver agreement. *Personnel Psychology* 16, 373-384

British Standards Institution. 1979. Precision of test methods I: Guide for the determination and repeatability for a standard test method. London: BSI.

Broekmans FJ, Kwee J, Hendriks DJ, et al. 2006 A systematic review of tests predicting ovarian reserve and IVF outcome. *Human Reproduction Update* 12, 685-718.

Christensen JT, Boldsen J, Westergaard JG. 1997 Ovarian volume in gynecologically healthy women using no contraception, or using IUD or oral contraception. *Acta Obstetrica et Gynecologica Scandinavica* 76, 784-78.

Frattarelli JL, Levi AJ, Miller BT, et al. 2004 Prognostic use of mean ovarian volume in in vitro fertilization cycles: a prospective assessment. *Fertility and Sterility* 82, 811-815.

Higgins RV, van Nagell JR, Woods CH, et al. 1990 Interobserver variation in ovarian measurements using transvaginal sonography. *Gynecologic Oncology* 39, 69-71.

Järvelä IY, Sladkevicius P, Tekay AH, et al. 2003 Intraobserver and interobserver variability of ovarian volume, gray-scale and color flow indices obtained using transvaginal three-dimensional power Doppler ultrasonography. *Ultrasound in Obstetrics and Gynecology* 21,277-82.

Jayaprakasan K, Hilwah N, Kendall NR, et al. 2007 Does 3D ultrasound offer any advantage in the pre-treatment assessment of ovarian reserve and prediction of outcome after assisted reproduction treatment? *Human Reproduction* 22, 1932-1941.

Kyei-Mensah A, Maconochie N, Zaidi J, et al. 1996a Transvaginal three-dimensional ultrasound: repeatability of ovarian and endometrial volume measurements. *Fertility and Sterility* 66, 718-722.

Kyei-Mensah A, Zaidi J, Pittrof R, et al. 1996b Transvaginal three-dimensional ultrasound: accuracy of follicular volume measurements. *Fertility and Sterility* 65, 371-376

Lass A, Silye R, Abrams DC, et al. 1997a Follicular density on ovarian biopsy of infertile woman: a novel method to assess ovarian reserve. *Human Reproduction* 12, 1028-1031.

Lass A, Skull J, McVeigh E, et al. 1997b Measurement of ovarian volume by transvaginal sonography prior to human menopausal gonadotrophin hyperstimulation can predict poor

response of infertile patients in an IVF programme. *Human Reproduction* 12, 294-297.

Mercé LT, Gómez B, Engels V, et al. 2005 Intraobserver and interobserver repeatability of ovarian volume, antral follicle count, and vascularity indices obtained with transvaginal 3-dimensional ultrasonography, power Doppler angiography, and the virtual organ computer-aided analysis imaging program. *Journal of Ultrasound in Medicine* 24, 1279-87.

Oppermann K, Fuchs SC, Spritzer PM. 2003 Ovarian volume in pre- and perimenopausal women: a population-based study. *Menopause*. 10, 209-213.

Raine-Fenning NJ, Campbell BK, Clewes JS, et al. 2003a The interobserver reliability of ovarian volume measurement is improved with three-dimensional ultrasound, but dependent upon technique. *Ultrasound in Medicine & Biology* 29, 1685-1690.

Raine-Fenning NJ, Clewes JS, Kendall NR et al. 2003b The interobserver reliability and validity of volume calculation from three-dimensional ultrasound datasets in the in vitro setting. *Ultrasound in Obstetrics and Gynecology* 21, 283-291.

Rovas L, Sladkevicius P, Strobel E, et al. 2005 Intraobserver and interobserver repeatability of three-dimensional gray-scale and power Doppler ultrasound examinations of the cervix in pregnant women. *Ultrasound in Obstetrics and Gynecology* 26, 132-7.

Saxton DW, Farquhar CM, Rae T, et al. 1990 Accuracy of ultrasound measurement of female pelvic organs. *British Journal of Obstetrics and Gynaecology* 97, 695-699.

Schild RL, Knobloch C, Dorn C, et al. 2001 The use of ovarian volume in an in vitro fertilization programme as assessed by 3D ultrasound. Archives of Gynecology & Obstetrics 265, 67-72.

Sharara FI, McClamrock HD. 1999 The effect of aging on ovarian volume measurements in infertile women. Obstetrics & Gynecology 94, 57-60.

Syrop CH, Dawson JD, Husman KJ, et al. 1999 Ovarian volume may predict assisted reproductive outcomes better than follicle stimulating hormone concentration on day 3. Human Reproduction. 14, 1752-1756.

Tomas C, Nuojuu-Huttunen S, Martikainen H. 1997 Pretreatment transvaginal ultrasound examination predicts ovarian responsiveness to gonadotrophins in in-vitro fertilization. Human Reproduction 12,220-223.

Valentin L, Bergelin I. 2002 Intra- and interobserver repeatability of ultrasound measurements of cervical length and width in the second and third trimesters of pregnancy. Ultrasound in Obstetrics and Gynecology 20, 256-62.

Wallace WH, Kelsey TW 2004 Ovarian reserve and reproductive age may be determined from measurement of ovarian volume by transvaginal sonography. Human Reproduction 19, 1612-1617.

Walter SD, Eliasziw M, Donner A. 1998 Sample size and optimal designs for reliability studies. Statistics in Medicine 17, 101-110.

Summary for lay readers

Measuring ovarian volume has been suggested as a possible screening test to assess a woman's ovarian reserve. In order for a screening tool based on ovarian volumes to be clinically useful, knowledge of the precision or reproducibility of the measurement technique used is essential. Recent advances in ultrasound scanning technique allow the measurement of volumes in three dimensions (3D) rather than the traditional estimation in 2 dimensions (2D). Transvaginal 2D and 3D ultrasound examinations were performed on 49 women attending a tertiary centre for investigation or treatment for subfertility between January and May 2006. Two observers calculated ovarian volume using both 2D techniques (prolate ellipsoid formula) and 3D techniques (virtual organ computer-aided analysis (VOCAL)) with rotation steps of 30 degrees (3D-30). For the four comparisons (inter- and intra-observer; 2D and 3D-30) we obtained intraclass coefficients of 0.97 to 0.98; and standard errors ranging from 17% to 14% (for inter-observer 2D and intra-observer 3D, respectively). The corresponding Coefficients of Repeatability ranged from 33% to 28%. Our results suggest that measurement of transvaginal ovarian volumes using both 2D and 3D ultrasound is imprecise for individuals. The imprecision is greater for lower ovarian volumes, which may be important in clinical practice. The average of two or more measurements is likely to be more accurate than a single measurement.

