Full title: **Self-reported transvaginal ultrasound visualization of normal ovaries in postmenopausal women is not reliable: results of expert review of archived images in UKCTOCS**

Short title: UKCTOCS TVS QC

Authors:

Will Stott[1], Stuart Campbell[2], Angelo Franchini[3], Oleg Blyuss[1], Alexey Zaikin[1], Andy Ryan[1], Chris Jones[1], Aleksandra Gentry-Maharaj[1], Gwendolen Fletcher[1], Jatinderpal Kalsi[1], Steve Skates[4], Max Parmar[5], Nazar Amso[6], Ian Jacobs[1,7], Usha Menon[1]

1: Women's Cancer, UCL EGA Institute for Women's Health, London, UK
2: Create Health Clinic, London, UK
3: London School of Hygiene and Tropical Medicine, London, UK
4: Biostatistics Centre, Massachusetts General Hospital, Boston, MA, USA
5: Medical Research Council Clinical Trials Unit at UCL, London, UK
6: School of Medicine, College of Biomedical and Life Sciences, Cardiff University, Cardiff
7: University of New South Wales, Australia

Corresponding author:

Usha Menon. Women's Cancer, UCL Institute for Women's Health, 1st Floor, Maple House, 149 Tottenham Court Road, London, W1T 7NF

u.menon@ucl.ac.uk (+44 203 447 2125)

Keywords: Ovarian Cancer Screening, UKCTOCS, Transvaginal Sonography (TVS), Ultrasound, Expert review, Automated Image Analysis, Quality Control (QC), Visualization Rate (VR).

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/uog.18836

ABSTRACT

Objective

In UKCTOCS self-reported visualization rates(srVR) at annual TVS scan was a key quality control(QC) metric. Our objective was to independently assess srVR using expert review and develop software capable of monitoring it.

Methods

Images from 1,000 examinations randomly selected from 68,951 archived annual TVS exams undertaken between 2008-2011 where the ovaries were reported as 'seen and normal' were reviewed by a single expert. Software was developed to identify exact images used to measure ovaries by measuring caliper dimensions and matching them to that recorded by the sonographer. A logistic regression classifier to determine visualization was trained and validated using ovarian dimension and visualization data reported by the expert .

Results

The expert confirmed both ovaries were visualized (cVR-Both) in 50.2%(502/1000) of the exams. The software identified the measurement image in 534 exams which were split 2:1:1 providing training, validating and testing data. Classifier accuracy on validation data was 70.9%(CI-95% 70.0,71.8). Analysis of test data (133 exams) resulted in sensitivity of 90.5%(CI-95% 80.9,95.8) and specificity of 47.5%(CI-95% 34.5,60.8) in detecting expert confirmed cVR-Both.

Conclusions

Our results suggest that in a significant proportion of TVS annual screens the sonographers may have mistaken other structures for normal ovaries. It is uncertain whether or not this affected the sensitivity and stage at detection of ovarian cancer in the ultrasound arm of UKCTOCS, but we conclude QC metrics based on self-reported visualization of normal ovaries are unreliable. The classifier shows some potential for addressing this problem, though further research is needed.

INTRODUCTION

Transvaginal ultrasound (TVS) is widely used for pelvic imaging both in the context of patient management and in ovarian cancer screening. Visualization of ovaries is a desired prerequisite but can be challenging in older women as the ovaries are typically shrunken or more difficult to locate.

Visualization rate (VR) is a widely used quality control (QC) metric for TVS scanning in the context of ovarian cancer screening.[1] It is the percentage of all exams performed by the sonographer in which the ovaries are identified. However, there is variation across different studies in terms of how VR is defined with some reporting visualization of both ovaries (VR-Both) and some of one or both ovaries (Table 1).[1-6] We believe that in the context of ovarian cancer screening, VR-both is the more meaningful metric as early cancer can begin in one ovary before spreading to the contralateral ovary.

Obtaining reliable VR data is challenging as ovarian visualization is subjective and sensitive to inter/intra observer variation.[7] VR is also problematic as all previous studies[1-6 8 9] have calculated ovarian VR using visualization data self-reported by the sonographer. In UKCTOCS self-reported VR was the QC metric used during annual ultrasound screening. However, static ultrasound images taken at the time of the exam were centrally archived so an apportunity was provided to retrospectively investigate whether ovarian visualization had been achieved. We are not aware of any previous study that has attempted such TVS validation apart from our group's audit of seven sonographers reporting high VR (Stott et al, *manuscript in preparation*).[10]

In this current study we report on a retrospective expert review of archived static images from a random sample of annual TVS examinations classified as normal and performed between 2008-2011 after accreditation had been introduced and quality monitoring had been improved.[2] Our study also attempts to address the problem of obtaining an objective measure of VR free from inter/intra observer variability by constructing a software classifier trained using data from the expert review with the aim of helping to drive future quality improvement in TVS.

METHODS

UKCTOCS is a multicentre randomized control trial involving 202,638 women volunteers from 13 trial centres (TC) in England, Wales and Northern Ireland. Inclusion criteria were postmenopausal women aged between 50-74 years at recruitment. Women with previous ovarian malignancy, bilateral oophorectomy, active non-ovarian malignancy, increased risk of familial ovarian cancer or participants in other ovarian cancer screening trials were excluded. The women were randomised into three groups: 1) ultrasound screening; n= 50,639, 2) multimodal screening using CA125 interpreted by the Risk of Ovarian Cancer (ROC) algorithm; n=50,640, 3) no screening (control); n=101,359.[11] Those in the ultrasound group underwent annual screening using TVS or a transabdominal (TAS) scan when a TVS was not acceptable to the volunteer. The details of the ultrasound screening process and its reporting have been previously described.[2] An important part of the exam results were capturing the dimensions of each ovary in two orthogonal planes which allowed ovary volume to be calculated. Annual screening in the ultrasound arm of UKCTOCS occurred

between 4th July 2001 and 21st December 2011. 48,250 volunteers received a total of 328,867 annual ultrasound screens of which 300,027 were TVS exams. A bespoke Trial Management System (TMS) implemented the algorithm described in the trial protocol for categorising TVS exams as abnormal, unsatisfactory or normal based on data reported by the sonographer for each ovary (Figure 2). This data included measurements (D1, D2, D3) for left (LO) and right ovary (RO) copied from values displayed by the ultrasound machine after the sonographer had placed calliper marks on the boundaries of the ovary in a static image captured for this purpose. A further bespoke computer system called the Ultrasound Record Archive (URA) was developed to archive these static images as reported elsewhere.[12]

The images from 216,152 TVS exams (72% of all TVS annual scans performed by UKCTOCS) were archived in the URA of which 113,092 exams were performed after January 2008 when quality monitoring had been improved, accreditation had been introduced and the ultrasound machine at all 13 trial centres were upgraded to Accuvix XQ model (Medison, Seoul, South Korea).[2] These later exams were performed by 141 sonographers all accredited to perform annual (level 1) TVS exams.[1]

The archived images from the random selection of 1,000 normal exams were used for the study dataset. The inclusion criteria for the exams were: a) annual TVS exams of women in the ultrasound screening group; b) images stored in the URA; c) performed after 1st January 2008; d) both ovaries measured / visualized; e) both ovaries categorised as having normal morphology. TAS only exams and those categorized as abnormal or unsatisfactory were excluded.

Expert Review

Images associated with each of the 1000 exams in the dataset were copied from the URA as 640 x 480 pixel greyscale bitmap files. A spreadsheet was prepared containing hyperlinks to these images so the reviewer could display them by selecting the appropriate cell, as reported elsewhere.[12] Bespoke software was used to process the images in order to measure the calliper marks. The resultant dimensions were matched against the ovary dimensions recorded in the TMS in order to identify the exact image the sonographer had used to measure the ovary.[12] The spreadsheet was annotated to indicate images that had been used to measure ovaries. However, the expert reviewed all the images for each exam in the dataset to detect any bias arising from software selection.

A single expert in gynaecological scanning reviewed the images for each exam and recorded assessments of left and right ovary using one of the following categorical variables; visualised and correctly measured, visualised but poorly measured, not visualised, other images which were not of the adnexal region such as uterus were marked as not appropriate. Criteria used to indicate that an ovary was not visualised (Figure 1) were an irregular or indistinct outline, a heterogeneous echogenicity of the stroma and an outline that could be be identified as part of a larger shape which was usually bowel. In practical terms the expert mentally removed the callipers and if the resulting shape did not resemble an ovary then the image was classed as “not visualised”. This was occasionally confirmed by measurements which were clearly outside the normal range expected for a postmenopausal ovary.

Construction of Logistic Regression Classifier

Statistical learning techniques were used to construct a logistic regression classifier using R v3.3.2. It used the train() function provided by CARET (Classification and REgression Training) package v6.0.71 with the generalised linear model (glm) specified as its method parameter. Ovary dimensions and ovary type (left or right) were used as candidate feature data. Ovary visualization (true or false, as judged by the expert) was used as the target value.

Statistical Analysis

Visualization rates were calculated from results of the expert review of all the images associated with the 1,000 exams in the dataset. An ovary was defined as 'seen' when the expert reviewer categorised the image using any of the categorical variables 'visualized and correctly measured' or 'visualized but poorly measured'. The use of any other categorical variables was defined as 'not seen'. Visualization rates were calculated for all 1,000 exams in the dataset using various VR definitions (Table 2) .

The exams in the dataset were processed to create two further subsets: 1) the 'match' subset containing examinations for which software found the exact images used by the sonographer to measure the left and right ovary. The software was set to only identify exams where images of the transverse and longitudinal planes of the ovary were saved in the "split-screen" function of the ultrasound machine as one image for each ovary and the software caliper measurements matched that reported by the sonographer. To facilitate analysis examinations which did not have both longitudinal section (LS) and transverse

section (TS) for both left and right ovary (TSLS-Both) were excluded. 2) The 'no match' subset containing the examinations for which software could not find the exact images used by the sonographer to measure the left and right ovary. Visualization Rates (VR) were calculated for the exams in both subsets using the definitions in Table 2 so that the differences between them could be assessed.

The 'match' subset was randomly split in the ratio 2:1:1 (training, validation, test) in order to build the logistic regression classifier. Various combinations of features were used to constuct models from the same selection of training and validation data so the performance of each could be evaluated in terms of accuracy. The combination of features that offered best performance was selected and the data was randomly split using different seed values so that performance could be measured for different (same sized) collections of training and validation data as randomly selected. The following were calculated: Mean and 95% confidence intervals for accuracy (mean true positives plus mean true negatives divided by total observations for ovary dimensions in the randomly selected validation data); Sensitivity (mean true positives divided by sum of mean true positives and mean false negatives); and Specificity (mean true negatives divided by sum of means true negatives and mean false positives). Mean values for true positive, true negative, false positive, false negatives (as defined in Table 3) were obtained by averaging values obtained for each selection of exams used as validation data as randomly split from the 'match' data subset. In this way the classifer performance metrics were not dependent on any particular selection of exams from the 'match' data subset.

The combination of features with the best performance was taken from exams in the test data and applied to the classifier. The results were recorded as classifier visualization; ovary visualized, or not. These results were used to determine visualization for each exam and these values were used to calculate the cVR-Both for the test data.

RESULTS

Annual TVS examinations with archived images performed after 1st January 2008 were categorised by the TMS according to the trial protocol (Figure 2) as normal (105,176; 93%), abnormal (5,097;4.5%) and unsatisfactory (2,820; 2.5%); total 113,093 exams. The dataset of 1,000 exams was randomly selected from 68,931 of the 105,176 normal exams that had both ovaries reported as 'seen' (visualized). This dataset had 4,654 images; mean 4.6 images per exam and range 1-15.

The results of the review by the one expert of images from 1,000 TVS exams allowed calculation of VR, but the values changed significantly depending on the definition used for visualization. Using a definition of both ovaries visualized (cVR-Both) the value of VR is 50.2%, but the VR value changes to 79.2% if a definition of one or both ovaries is applied (Table 2).

The software was set to only identify exams where the transverse and longitudinal sections were in the same image for both ovaries and the software caliper measurements matched that reported by the sonographer. This was possible in a 'subset' of 534 exams (53.4%). The images used to measure the ovaries were identified in a further 17 (1.7%) exams but were excluded by the

software as there were multiple images. In the remaining 449 exams (44.9%) the images used to measure ovaries could not be recovered for the following reasons - 8.6% duplicate images, 16.4% unresolved, 3.9% process failure, 16.0% images with non standard caliper marks.[12] The expert VR results of images in this 534 exam 'match' subset tended to be higher than that of images in the remaining 466 exams 'no-match' subset although the difference in rates was not significant (Supplementary Table 1).

The classifier's performance was evaluated using the validation data; 268 of 1068 ovary dimensions in the 'match' subset of 534 exams. Thirty different collections of validation data were generated by randomly splitting the subset using different seed values, each having feature and target data from left and right ovary in 134 exams; 268 total. The results of each collection were calculated as described in methods; mean accuracy 70.9%(CI-95% 70.0,71.8), mean sensitivity 93.0%(CI-95% 92.1,94.0), mean specificity 27.3%(CI-95% 25.4,29.3). The Receiver-Operator Curve (ROC) shown in Figure 3 was produced from validation data in the same random split of the 'match' data subset which contained the test data used to calculate cVR-Both.

The test data was formed by 266 ovarian dimensions from 133 exams in the 'match' data subset which had not been used for training or validation. Fourty seven sonographers performed these exams with the number of exams by individual sonographers having a range of 1-15 and mean of 2.83. When the test data was applied to the classifier, cVR-Both was 73.7% compared with the gold standard of 55.6% found by the expert. The expected accuracy was calculated as 61.8% which gave a Kappa value of 0.24 (judged only fair

according to Landis Koch interpretation)[13] with sensitivity 90.5% (CI-95% 80.9,95.8) and specificity 47.5% (CI-95% 34.5,60.8) as shown in Table 3.

DISCUSSION

A retrospective review by a single expert of archived images from a random selection of 1,000 annual UKCTOCS TVS exams where both ovaries were "seen and normal", demonstrated that the expert could definitely confirm both ovaries visualized in only half of the archived exams. In the remaining exams, the expert considered that the sonographer had mistaken some other structure for an ovary, most commonly bowel (Figure 1). As far as we are aware no other screening study has undertaken a similar independent review of normal ovarian scans. Our findings suggest that self-reported sonographer VR unless confirmed by independent review is not reliable as a QC metric and should be used with caution in the future.

It is generally accepted that the success of any screening programme for ovarian cancer using TVS is highly dependent on sonographers finding any small tumours that might exist in either ovary. Models[14] estimate that majority of high grade serous ovarian cancers progress to Stage III/IV at a median diameter of about 3 cm. Identifying half these tumors in Stage I/II at annual screen would require detection of tumors 1.3 cms in diameter but to achieve a 50% mortality reduction it would be necessary to detect tumors 0.5 cm in diameter. Identifying such small tumours is very challenging even for expert sonographers. Therefore different levels of sonographer skill and experience might explain variation in outcome between the single centre Kentucky Ovarian

Cancer Ultrasound Screening Study[6] and UKCTOCS as well as other large scale multicentre trials (e.g. PLCO) where it is not feasible for a small group of experts to deliver annual population screening.

We cannot assess the impact on stage shift of the discrepancy between level I ultrasonographers and the expert on ovarian visualization in the ultrasound arm of UKCTOCS.[15] However, all archived examinations preceding ovarian cancer diagnosis in the ultrasound arm of UKCTOCS were reviewed in the course of the trial and collation of these results should provide further insights.

A key quality metric in all ultrasound screening trials is self-reported VR. In UKCTOCS, a quality monitoring programme with regular feedback was in place throughout.[2] It included 6 monthly monitoring of self-reported VR together with other data such as ovarian size and missing/inaccurate information entered into the TMS. In addition UKCTOCS Level I sonographers with VR below 60% were subject to targeted training.[2] To what extent these measures might have resulted in some sonographers designating the ovary as 'seen' when in doubt is difficult to ascertain.

The use of statistical learning techniques to construct a logistic regression classifier raises the possibility of obtaining independent reliable QC metrics that can be applied at low cost to large scale TVS examinations. We report on a classifier using ovarian dimensions to identify the ovary. However, specificity was low. It is possible that better performance would have been achieved had morphological features been included as well as ovary dimensions.

Strengths and Limitations

We are not aware of any other study which has reported on a similar independent review of TVS examinations of archived normal TVS ovarian examinations. Key strengths include the scale in terms of number of examinations and sonographers from multiple centres reflecting the reality of a population ultrasound screening programme; archived images available for 72% of all TVS annual scans performed, random selection of examinations from those classified as normal and use of the exact image that was used to measure the ovary. A limitation was the stringent critieria used by the software to identify images that limited the number of images that could be assessed by the QC classifier. In prospective studies, this could be addressed by the ultrasonographer 'flagging' the exact ovarian images during scanning.

A major limitations were that the review was performed by only one expert. Given the known subjectivity of TVS, more robust estimates of expert VR would have been obtained by repeat assessment of random subsets to assess both intra and inter observer variability. In an audit of seven sonographers reporting high sVR in UKCTOCS, there was significant variation in inter-observer agreement between six experts (Stott et al, *manuscript in preparation*).[10]

Other Studies

The above audit performed in 2009 involved a similar review of the images used to measure ovaries from TVS exams performed by UKCTOCS sonographers. In this study two teams of three experts agreed that visualization of both ovaries could not be confirmed in a proportion of exams which had been reported as normal However, further conclusions about UKCTOCS scanning quality could not be made due to the small number of sonographers audited (7) and the way they were selected. A report about this audit is currently being prepared for publication.

Conclusion

Our results suggest that reliable quality control for TVS cannot be achieved using sonographers self-reporting ovary visualization because in almost half the annual TVS examinations performed by UKCTOCS after Jan 2008 both ovaries were not visualised. The results highlight the subjective nature of grey scale ultrasound imaging and the role of operator experience in scanning older postmenopausal women. It is uncertain whether or not this affected the sensitivity and stage at detection in the ultrasound arm of UKCTOCS. However, this study does underline the challenges of delivering large-scale TVS screening for ovarian cancer and the need to base its quality management on independent as well as objective quality control metrics. In this regard the classifier produced for our study shows some potential, though further research is needed before it could be used in a TVS Quality Improvement (QI) programme.

ACKNOWLEDGMENTS

We are very grateful to the many volunteers throughout the UK who participated in the trial and the entire medical, nursing, administrative staff and sonographers who work on the UKCTOCS. In particular, the UKCTOCS Centre leads : Keith Godfrey, Northern Gynaecological Oncology Centre, Queen Elizabeth Hospital, Gateshead; David Oram, Department of Gynaecological Oncology, St. Bartholomew's Hospital, London, Jonathan Herod, Department of Gynaecology, Liverpool Women's Hospital, Liverpool, Karin Williamson, Department of Gynaecological Oncology, Nottingham City Hospital Nottingham; Howard Jenkins, Department of Gynaecological Oncology, Royal Derby Hospital, Derby; Tim Mould, Department of Gynaecology, Royal Free Hospital; Robert Woolas, Department of Gynaecological Oncology, St. Mary's Hospital, Portsmouth; John Murdoch Department of Gynaecological Oncology, St. Michael's Hospital, Bristol; Stephen Dobbs Department of Gynaecological Oncology, Belfast City Hospital, Belfast; Simon Leeson Department of Gynaecological Oncology, Llandudno Hospital, North Wales; Derek Cruickshank, Department of Gynaecological Oncology, James Cook University Hospital, Middlesbrough.

Disclosure of Interests

UM has stock ownership and research funding from Abcodia. She has received grants from Medical Research Council (MRC), Cancer Research UK (CR UK), the National Institute for Health Research (NIHR), and The Eve Appeal (TEA). IJJ reports personal fees from and stock ownership in Abcodia as the non-executive director and consultant. IJJ reports personal fees from Women's Health Specialists as the director. IJJ has a patent for the Risk of Ovarian Cancer algorithm and an institutional licence to Abcodia with royalty agreement. He is a trustee (2012–14) and Emeritus Trustee (2015 to present) for The Eve Appeal. IJJ has received grants from the MRC, CR UK, NIHR, and TEA. The remaining authors declare no competing interests.

Contribution to Authorship

UM, SC, AR, WS and AGM were responsible for the design of the expert review. SC performed the expert review of the images. WS analysed the images using the software he had developed and collated all the data. WS, CJ, OB and AF undertook the data analysis. WS, AF, OB, CJ, SC and UM were involved in data interpretation. WS and UM drafted the paper. All contributed to editing the paper and approved the final version.

Ethics approval

The UKCTOCS study was approved by North West Multicentre Research Ethics Committee 21/6/2000; MREC reference 00/8/34. It is registered as an International Standard Randomised Controlled Trial (no. ISRCTN22488978).

Funding

The UKCTOCS trial was core funded by the Medical Research Council, Cancer Research UK, and the Department of Health with additional support from the Eve Appeal, Special Trustees of Bart's and the London, and Special Trustees of UCLH. The researchers at UCL were supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. The researchers are independent from the funders.

References

1. Sharma A, Burnell M, Gentry-Maharaj A, Campbell S, Amso NN, Seif MW, Fletcher G, Brunel C, Turner G, Rangar R, Ryan A, Jacobs I, Menon U, United Kingdom Collaborative Trial of Ovarian Cancer S. Factors affecting visualization of postmenopausal ovaries: descriptive study from the multicenter United Kingdom Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology 2013;**42**(4):472-7.
2. Sharma A, Burnell M, Gentry-Maharaj A, Campbell S, Amso NN, Seif MW, Fletcher G, Brunell C, Turner G, Rangar R, Ryan A, Jacobs I, Menon U. Quality assurance and its impact on ovarian visualization rates in the multicenter United Kingdom Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology 2016;**47**(2):228-35.
3. Bodelon C, Pfeiffer RM, Buys SS, Black A, Sherman ME. Analysis of serial ovarian volume measurements and incidence of ovarian cancer: implications for pathogenesis. Journal of the National Cancer Institute 2014;**106**(10).
4. van Nagell JR, Jr., DePriest PD, Reedy MB, Gallion HH, Ueland FR, Pavlik EJ, Kryscio RJ. The efficacy of transvaginal sonographic screening in asymptomatic women at risk for ovarian cancer. Gynecologic oncology 2000;**77**(3):350-6.
5. van Nagell JR, Jr., DePriest PD, Ueland FR, DeSimone CP, Cooper AL, McDonald JM, Pavlik EJ, Kryscio RJ. Ovarian cancer screening with annual transvaginal sonography: findings of 25,000 women screened. Cancer 2007;**109**(9):1887-96.
6. van Nagell JR, Jr., Miller RW, DeSimone CP, Ueland FR, Podzielinski I, Goodrich ST, Elder JW, Huang B, Kryscio RJ, Pavlik EJ. Long-term survival of women with epithelial ovarian cancer detected by ultrasonographic screening. Obstet Gynecol 2011;**118**(6):1212-21.
7. Higgins RV, van Nagell JR, Jr., Woods CH, Thompson EA, Kryscio RJ. Interobserver variation in ovarian measurements using transvaginal sonography. Gynecologic oncology 1990;**39**(1):69-71.
8. Ludovisi M, Mavrelos D, Jurkovic D. Re: factors affecting visualization of postmenopausal ovaries: descriptive study from the multicenter United Kingdom Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). Ultrasound in obstetrics & gynecology : the official journal of the

International Society of Ultrasound in Obstetrics and Gynecology 2014;**43**(5):600.
9. Gollub EL, Westhoff C, Timor-Tritsch IE. Detection of ovaries by transvaginal sonography in postmenopausal women. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology 1993;**3**(6):422-5.
10. Stott W, Ryan A, Gentry-Maharaj A, Fletcher G, Burnell M, Amso N, Seif M, Ferguson A, Turner G, Brunell C, Ford K, Rangar R, Jones C, Jacobs I, Parmar M, Menon U, Campbell S. Audit of transvaginal sonography (TVS) by sonographers from the United Kingdom Collaborative Trial of Ovarian Cancer Screening (UKCTOCS) (manuscript in preparation).
11. Menon U, Gentry-Maharaj A, Ryan A, Sharma A, Burnell M, Hallett R, Lewis S, Lopez A, Godfrey K, Oram D, Herod J, Williamson K, Seif M, Scott I, Mould T, Woolas R, Murdoch J, Dobbs S, Amso N, Leeson S, Cruickshank D, McGuire A, Campbell S, Fallowfield L, Skates S, Parmar M, Jacobs I. Recruitment to multicentre trials--lessons from UKCTOCS: descriptive study. Bmj 2008;**337**:a2079.
12. Stott WPQ. Use of Software Tools to Implement Quality Control of Ultrasound Images in a Large Clinical Trial [PhD]. University College London, 2016.
13. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;**33**(1):159-74.
14. Brown PO, Palmer C. The preclinical natural history of serous ovarian cancer: defining the target for early detection. PLoS medicine 2009;**6**(7):e1000114.
15. Jacobs IJ, Menon U, Ryan A, Gentry-Maharaj A, Burnell M, Kalsi JK, Amso NN, Apostolidou S, Benjamin E, Cruickshank D, Crump DN, Davies SK, Dawnay A, Dobbs S, Fletcher G, Ford J, Godfrey K, Gunu R, Habib M, Hallett R, Herod J, Jenkins H, Karpinskyj C, Leeson S, Lewis SJ, Liston WR, Lopes A, Mould T, Murdoch J, Oram D, Rabideau DJ, Reynolds K, Scott I, Seif MW, Sharma A, Singh N, Taylor J, Warburton F, Widschwendter M, Williamson K, Woolas R, Fallowfield L, McGuire AJ, Campbell S, Parmar M, Skates SJ. Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. Lancet 2016;**387**(10022):945-56.

FIGURES

Figure 1

Figure 1: Samples of images created by sonographers for measuring ovaries. The expert judged the images of left and right ovary (left) as normal and correctly measured by the sonographer. However, in the case of the other image (right) the expert considered the sonographer had mistakenly measured a section of bowel rather than the left ovary indicated by the annotation as the haustrations of large bowel are clearly visible in the structure marked by the callipers

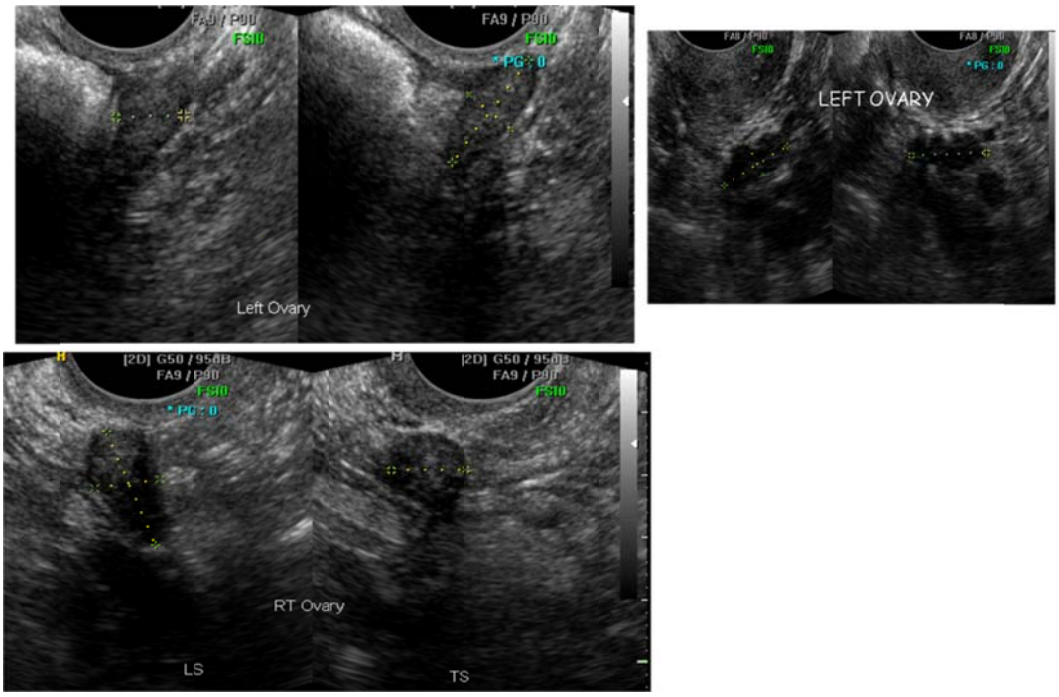


Figure 2 (same as previously - included here for completeness)

Figure 2: UKCTOCS algorithm for classification of pelvic scans based on data reported by the sonographer

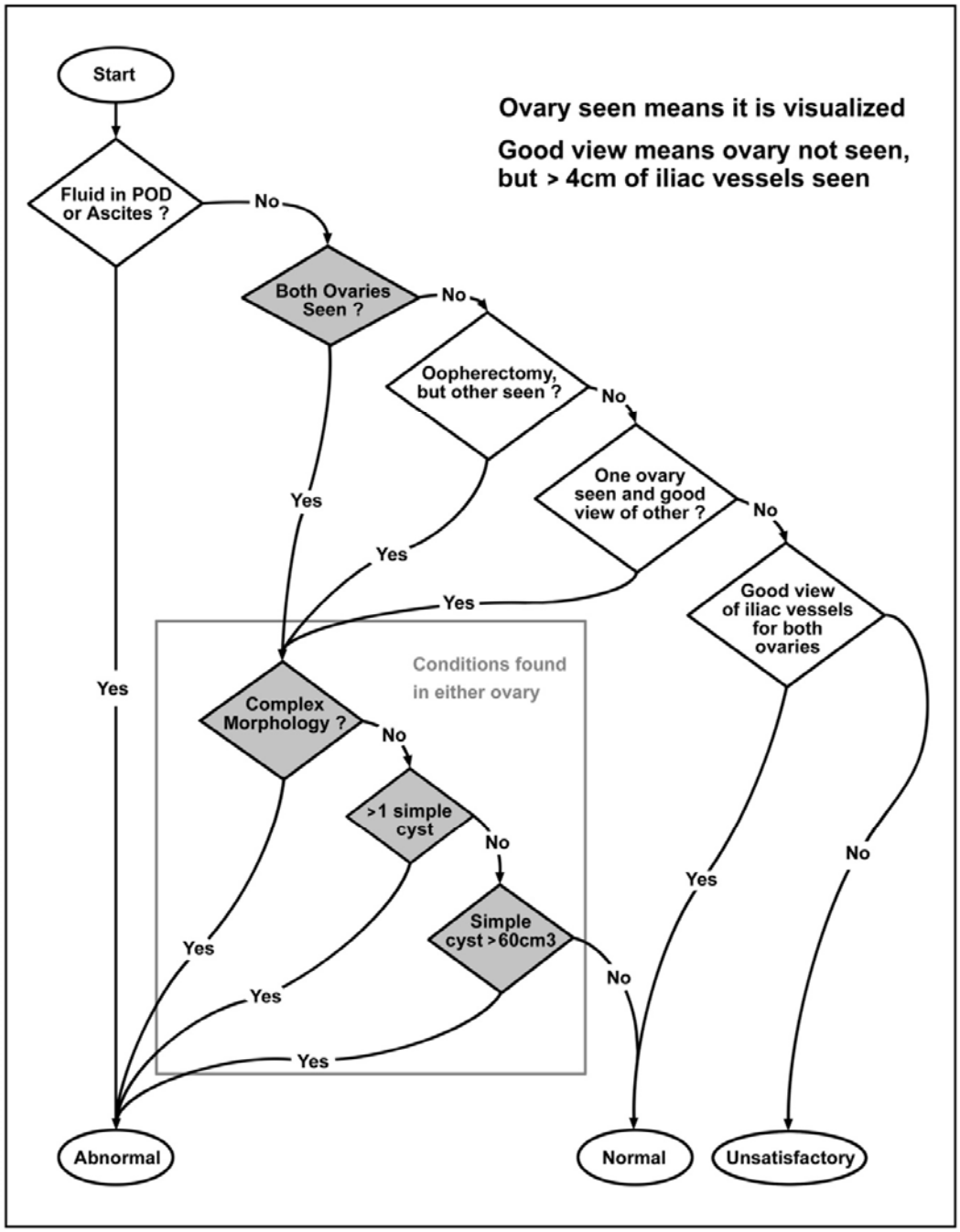
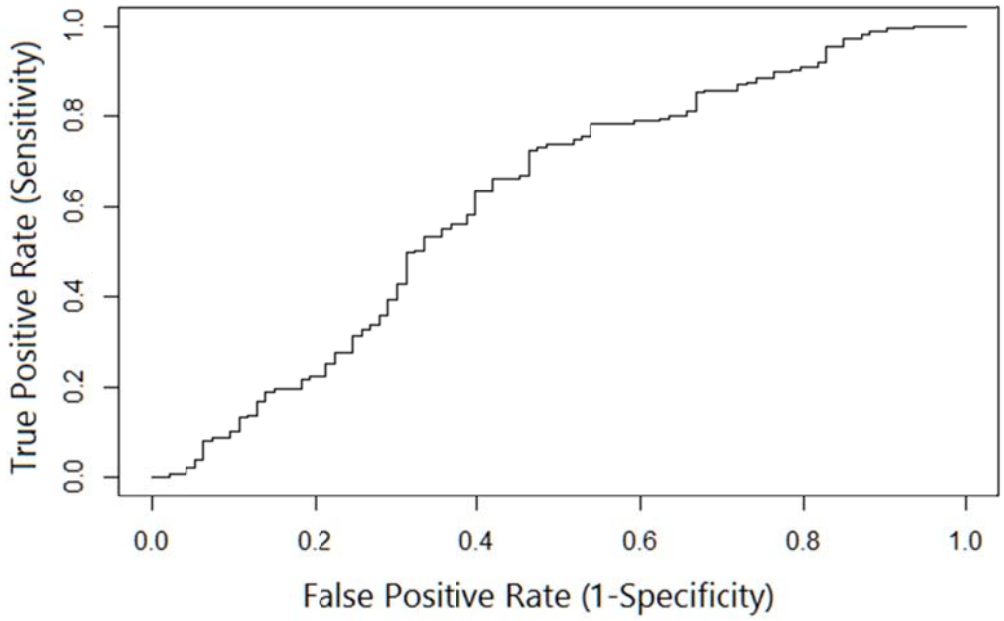


Figure 3

Figure 3: Receiver-Operator Curve (ROC) showing classifier performance for individual ovaries using validation data values from the same random split of the 534 exam 'match' data subset into training, validation and test data



TABLES

Table 1: Variation in ovarian Visualization Rate (VR) on TVS related to different definitions

Table 2: Variation in expert assessed Visualization Rates using different VR definitions

Table 3: Contingency table comparing Classifier to Expert visualization of both ovaries (cVR-Both) in the test data set

Supplementary Table 1: Visualization Rates (VR) from expert review for 'match' and 'no match' subsets of the study dataset categorised by visualization definition given in Table 2. The 'match' subset contains exams for which the exact images used to measure left and right ovary can be identified by the software and the 'no match' subset contains the remaining exams

Trial or study	Exam numbers	Dates	Definition of Visualization Rate (VR)	Reported Visualization Rate (VR)
UKCTOCS[2]	270,035	June 2001-Dec 2011	RO or both	72.7%
			One or both	84.5%
UKCTOCS[1]	43,867	June 2001-Aug 2007	RO or both	66.8%
			LO or both	65.5%
PLCO[3]	102,787	1993-2009	Both	60%
Kentucky[4]	57,214	1987-1999	One or both	79.2%
Kentucky[5]	120,569	1987-2005	One or both	84%
Kentucky[6]	205,190	1987-2011	One or both	87.6%
Ludovisi et al study[8]	6649	Oct 2008-Sept 2013	RO or both	84.1%
			LO or both	82.4%
Gollub et al study[9]	206	June 1988-Mar1989	Both	49%
			One or both	80%

Table 1: Variation in Visualization Rate (VR) in reports of TVS scanning related to different definitions of ovary visualization.

Expert VR Definition	Exams with ovarian images identified by software		Exams with ovarian images not identified by software		All exams	
	Count (n=534)	**Expert VR %**	**Count (n=466)**	**Expert VR %**	**Count (n=1000)**	**Expert VR %**
RO or Both right ovary or both ovaries seen	366	68.5	298	64.0	663	66.3
LO or Both left ovary or both ovaries seen	344	64.4	286	61.4	630	63.0
One or Both left or right ovary seen or both ovaries seen	430	80.5	362	77.7	792	79.2
Both (cVR-Both) both ovaries seen	280	52.4	222	47.6	502	50.2

Table 2: Variation in expert assessed Visualization Rates using different VR definitions.

		Visualization on expert review		
		Both ovaries visualized	**One or both ovaries not visualized**	**Total Exams**
Visualization by Classifier	**Both ovaries visualized**	67 (True positives)	31 (False positives)	98
	One or both ovaries not visualized	7(False negatives)	28 (True negatives)	35
Total Exams		74	59	133

Table 3: Contingency table comparing Classifier to Expert visualization of both ovaries (cVR-Both).