Published in final edited form as:

Obstet Gynecol. 2012 July ; 120(1): 104–112. doi:10.1097/AOG.0b013e31825bc6cf.

Interrater and Intrarater Reliability in the Diagnosis and Staging of Endometriosis

Karen C. Schliep, PhD, MSPH^a, Joseph B. Stanford, MD, MSPH^a, Zhen Chen, PhD^b, Bo Zhang, PhD^b, Jessie K. Dorais, MD^c, Erica Boiman Johnstone, MD^c, Ahmad O. Hammoud, MD^c, Michael W. Varner, MD^d, Germaine M. Buck Louis, PhD^b, C. Matthew Peterson, MD^c on behalf of the ENDO Study Working Group

^aDepartment of Family and Preventive Medicine, University of Utah, Salt Lake City, Utah;

^bDivision of Epidemiology, Statistics and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, the National Institutes of Health, Rockville, Maryland;

^cDivision of Reproductive Endocrinology and Infertility, Department of Obstetrics and Gynecology, University of Utah, Salt Lake City, Utah;

^dDivision of Maternal Fetal Medicine, Department of Obstetrics and Gynecology, University of Utah, Salt Lake City, Utah

Abstract

Objective: To estimate the interrater and intrarater reliability of endometriosis diagnosis and severity of disease among gynecological surgeons viewing operative digital images.

METHODS: Study population comprised a random sample (n=148; 36%) of women who participated in the Endometriosis: Natural History, Diagnosis and Outcomes Study. Four academic expert and four local, specialized expert surgeons reviewed the images, diagnosed the presence or absence of endometriosis for each woman, and rated severity using the revised American Society for Reproductive Medicine (ASRM) criteria. Interrater and intrarater level agreement were calculated for both endometriosis diagnosis and staging.

Results: The interrater reliability for endometriosis diagnosis among the eight surgeons was substantial: Fleiss $\kappa = 0.69$, 95% CI: 0.64–0.74. Surgeons agreed on revised ASRM endometriosis staging criteria after experienced assessment in a majority of cases (mean=61%, range: 52–75%) with moderate interrater reliability: Fleiss $\kappa = 0.44$, 95% CI: 0.41–0.47. The intrarater reliability for experienced assessment compared with computer-assisted revised ASRM staging was almost perfect (mean weighted $\kappa=0.95$, range: 0.89–0.99).

Corresponding Author: C Matthew Peterson, Division of Reproductive Endocrinology and Infertility, Department of Obstetrics and Gynecology, 30 North 1900 East, Suite 2B200, University of Utah, Salt Lake City, Utah 84132. Phone: 801-587-8303. Fax: 801-585-9295; c.matthew.peterson@hsc.utah.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Financial Disclosure: The authors did not report any potential conflicts of interest.

Conclusions: Substantial reliability was found for revised ASRM endometriosis diagnosis, while moderate reliability was observed for staging. Almost perfect reliability was observed for surgeons' rating of disease severity compared to computerized-assisted, checklist-based staging. Findings suggest that reliability in endometriosis diagnosis is not greatly altered by location or composition of surgeons, supporting the conduct of multisite studies or compilation of endometriosis data across clinical centers. While surgeons appear to be skilled at assessing endometriosis stage intuitively, how staging of disease burden correlates with clinical outcomes remains to be developed.

Précis:

Experienced gynecologic surgeons have substantial agreement about the revised American Society for Reproductive Medicine (ASRM) endometriosis diagnosis ($\kappa = 0.69$) and moderate agreement on the revised ASRM endometriosis staging ($\kappa = 0.44$).

INTRODUCTION

Surgical visualization for the diagnosis and staging of endometriosis, the current gold standard,^{1,2} is often recorded for later clinical, research, or medicolegal review.³ Gynecologists are often asked to review operative images to make judgments on treatment options for endometriosis. Consistency within and between evaluators is critical to the interpretation and application of findings. The opportunity for misclassification bias of endometriosis status among gynecologists threatens our ability to better understand the etiology and clinical management of the disease.

While endometriosis classification systems have been used for decades,⁴⁻⁶ few studies have evaluated the interrater and intrarater reliability of endometriosis diagnosis based on the revised American Fertility Society (rAFS) criteria,⁷⁻⁹ or the revised American Society for Reproductive Medicine (ASRM) criteria.¹⁰ Limitations of prior studies include restricting the population under review to women previously diagnosed with endometriosis,^{8,9} small numbers of assessments per rater,^{7,8,10} and restricting assessors to either one hospital⁷⁻⁹ or non-experts.¹⁰

Variability in the prevalence of diagnosed endometriosis in different clinical samples or study populations including its presence in asymptomatic women^{11,12} emphasizes the importance of a reliable assessment blind to prior clinical history and conducted on a population-based sample. Conducting such an assessment is important for evaluating the prudence of multi-site studies or the compilation of endometriosis data across clinical centers and also for understanding the adequacy of current staging systems. The primary purpose of this study is to estimate the interrater and intrarater reliability of the diagnosis and staging of endometriosis among experienced gynecologic surgeons practicing in a variety of clinical centers, after viewing operative digital images for the women participating in the Endometriosis: Natural History, Diagnosis and Outcomes (ENDO) Study.¹²

MATERIALS AND METHODS

A random sample of women was selected for study from the larger ENDO Study for the specific aim of conducting a reliability study. Briefly, the ENDO Study utilized a matched exposure cohort design to assess environmental chemicals and lifestyle behaviors associated with endometriosis in two study cohorts: operative and population.¹² ENDO Study operative participants comprised currently menstruating women, aged 18–44 years, undergoing laparoscopy or laparotomy (irrespective of indication) at one of 14 surgical sites in California or Utah, 2007–2009. Women with prevalent disease were excluded.

For purposes of this reliability study, we restricted to the Utah research sites, given that these enrolled 87% of study participants, and relied upon the operative cohort where the gold standard of disease visualization could be clinically determined.^{1,2} Laparoscopy alone is the current standard for the diagnosis of endometriosis and, by default, the visually defined revised ASRM staging.^{13–15} A random sample of women (n=148; 36%) was selected after stratifying on a postoperative diagnosis of endometriosis (yes/no). Of the 111 women with a postoperative diagnosis of endometriosis, 72 (65%) underwent a histologic evaluation for which 44 (61%) were histologically confirmed.

Surgeries were performed across various operating suites, with a variety of equipment available for recording video or still images. Operating surgeons were instructed to take intraoperative photos to document endometriosis or other gynecologic pathology using digital cameras attached to laparoscopes regardless of whether the woman was undergoing a laparoscopy or laparotomy. Specifically, surgeons were asked to photograph panoramic anterior and posterior views of the uterus and adnexal structures. Since there was a spectrum of image quality, ENDO staff and investigators reviewed all images for quality and rated the images as good or poor. Confirmation of image quality and selection was finalized by the Utah ENDO principal investigators (CMP, JBS).

Using a block randomization approach, we selected 148 women from the Utah operative cohort using the following stratification scheme: 105 women with a postoperative diagnosis of endometriosis and 43 women without a postoperative diagnosis of endometriosis. This stratification scheme was developed a priori to ensure the study had >99% and >85% statistical power for testing interrater reliability for the presence/absence of endometriosis and staging (between I/II and III/IV), respectively. Power calculations were based upon an alpha 0.05 and other assumptions as derived from the literature.⁷ Poor images (n=17) were intentionally included in our random sample to reflect a representative surgical cohort. Additionally, we did not wish to introduce selection bias by restricting our study sample to only women with good images, evidenced by the fact that within the entire Utah operative cohort (n=412), 52% of women with an endometriosis diagnosis had good quality images, while 21% of women without an endometriosis diagnosis had good quality images.

Four academic expert surgeons and four local, specialized expert (i.e., fellowship-trained) surgeons, determined a priori by ENDO principal investigators, were recruited for the study. Due to the extensive nature of the review, surgeons were not randomly selected for study recruitment. University affiliation was not a selection criterion. Academic expert surgeons

included physicians from a variety of North American centers who direct specialized training programs in laparoscopic gynecologic surgery, and who have extensive clinical and research experience in diagnosing and treating endometriosis. Local specialized expert surgeons included Utah physicians practicing in a variety of clinical centers with special surgical training and expertise in the diagnosis and treatment of endometriosis, and in the training of residents and fellows. The University of Utah Institutional Review Board approved this study and all physicians signed an informed consent document before being given access to the online review system. Raters were remunerated equally for their time and effort in completing ratings.

ENDO staff and investigators prepared anonymous digital images free of all clinical information using a standardized format designed to minimize rater fatigue and improve ease of use. The images were delivered to the clinical raters via an online system, which presented images one at a time per woman. The raters were asked to rate the image as poor, fair, good, or unable to assess. The raters were then asked to determine endometriosis status as follows: No endometriosis observed, Stage I (minimal), Stage II (mild), Stage III (moderate), Stage IV (severe), or indeterminate if unable to diagnose with reasonable accuracy. Prior to beginning their ratings, participating surgeons were asked to review the revised ASRM criteria for the staging of endometriosis.⁶ If the rater determined that endometriosis was present, they were also asked to complete a checklist for the following specific findings corresponding to the specific revised ASRM criteria: location of lesions, size of lesions, status of the posterior cul de sac, and the location and types of adhesions (with an option for not applicable). To avoid viewer fatigue, the online review system did not allow more than 90 minutes at a sitting.

Three outcomes were derived from each expert's rating: 1) a binary indicator of whether the rater reported endometriosis as present or absent, 2) the rater's categorization of endometriosis staging, 3) and the computer-assisted staging based upon the expert's checklist of findings and the revised ASRM algorithm.⁶

Descriptive statistics were calculated to summarize rater characteristics by rater type, expert versus local gynecological surgeons. Inter-rater agreement for the clinical diagnosis of endometriosis was evaluated by Kappa statistics (κ).^{16,17} The κ statistic summarizes the rating data into a contingency table, quantifying the proportion of chance-correct agreement relative to the maximum possible proportion of agreement beyond chance. If the raters are in complete (perfect) agreement, then $\kappa = 1$. When κ equals 0, the agreement is no better than what would be obtained by chance alone. For our reliability analyses, we utilized Landis and Koch's¹⁸ guidelines for interpreting κ statistics: $\kappa < 0.00$ represented poor agreement; $0.00 \leq \kappa < 0.20$ indicated slight agreement; $0.21 \leq \kappa < 0.40$ denoted fair agreement; $0.41 \leq \kappa < 0.60$ characterized moderate agreement; $0.61 \leq \kappa < 0.80$ defined substantial agreement, and a value of $\kappa \geq 0.80$ equated to almost perfect agreement. For binary outcomes (e.g., presence/absence of endometriosis), we computed pair-wise agreement using Cohen's κ and multi-rater agreement using Fleiss' multi-rater κ . For ordinal outcomes (e.g., staging of endometriosis), we computed pair-wise agreement using Cohen's weighted κ with squared weights and multi-rater agreement using Fleiss' multi-rater κ . For continuous outcome (i.e., revised ASRM score), we constructed pair-wise Bland-Altman plots¹⁹ to visualize the

agreement between any two raters and to assess whether differences between reviewers varied in a systematic way over the range of revised ASRM scores. Point estimates of κ and their 95% confidence intervals (CI) were estimated. For each κ , the P -value was calculated from the one-sided Wald test with null hypothesis $\kappa = 0.40$ versus alternative hypothesis $\kappa = 0.75$. If a woman had an “indeterminate” diagnosis for endometriosis, then that woman was excluded from the final data analysis under the assumption of data being missing at random. Analyses were performed in R version 2.13.1 (<http://www.r-project.org>) and SAS version 9.2 (SAS Institute, Cary, NC).

RESULTS

Eight raters (four academic expert and four local, specialized expert gynecological surgeons) assessed images during the summer of 2010. While fellowship trainings were similar for both groups of raters, the academic expert surgeons had practiced nearly three times the number of years since their fellowship (median 15.0, IQR 15.0–21.0, range 15–21 years) compared to the local, specialized expert surgeons (median 5.0, IQR 3.5–12.5, range 2–20 years). Academic expert surgeons also had more experience authoring, teaching, and serving as principal or co-investigator of funded studies addressing endometriosis compared to local, specialized expert surgeons. Neither the number of endometriosis patients seen per week nor the number of laparoscopies performed per month substantially differed between the rater groups. While there were no significant differences between groups in regards to age, race, marital status, income, and primary reason for surgery, women with an endometriosis diagnosis were less likely to previously been pregnant ($P=0.002$) or have had a live birth ($P<0.001$) compared to woman without an endometriosis diagnosis (Table 1). Among the 148 women, 121 (82%) had only digital photos, 3 (2%) had only digital video images, and 24 (16%) had both digital photos and video images. The eight raters determined that among the 148 women with images, 38% of the women had good operative images, 40% had fair operative images, 17% had poor operative images, and 6% had images unable to be assessed.

Endometriosis diagnosis was reported for approximately 66% of women, with little variation by type of rater, i.e., 65% among academic expert and 66% among local, specialized expert surgeons. In contrast, the distribution of revised ASRM severity varied by type of rater. Specifically, academic expert surgeons rated a lower incidence of stage II – IV (58% as stage I, 24% as stage II, 10% as stage III, and 8% as stage IV) compared with local, specialized expert surgeons (47% as stage I, 28% as stage II, 16% as stage III, and 10% as stage IV).

As can be seen in Table 2, the interrater reliability for the diagnosis of endometriosis (present or absent) among the raters based on digital images ranged from 0.47 to 0.86 with an overall Fleiss’s multi-rater κ of 0.69 (95% CI 0.64–0.74). The academic expert surgeons had substantial interrater reliability (Fleiss’s $\kappa=0.79$, 95% CI 0.70–0.88) compared with the local, specialized expert surgeons who had moderate agreement (Fleiss $\kappa=0.58$, 95% CI 0.50–0.66).

Surgeons agreed on revised ASRM endometriosis staging criteria in a majority of cases (mean=61%, range: 52–75%) with moderate interrater reliability when based on experienced

assessment (Fleiss $\kappa=0.44$, 95% CI 0.41–0.47, range 0.58–0.83) (Table 3) or when derived from the computer-assisted revised ASRM algorithm based on the reviewers' checklist of findings (Fleiss $\kappa=0.45$, 95% CI 0.42–0.48, range 0.55–0.80] (Table 4). The academic expert surgeons had moderate agreement (Fleiss $\kappa=0.44$, 95% CI 0.39–0.49] for staging of endometriosis after experienced assessment as did the local, specialized expert surgeons (Fleiss $\kappa=0.39$, 95% CI 0.34–0.43). A similar, albeit slightly higher, pattern was observed for computer-assisted staging for both groups of raters as evident by completely overlapping CIs (i.e., expert surgeons had $\kappa=0.46$; 95% CI 0.40–0.51 and local surgeons had $\kappa=0.45$; 95% CI: 0.40–0.51). The vast majority of interrater pairwise comparisons reached statistical significance for a preset hypothesis of $\kappa=0.75$ versus the null hypothesis of $\kappa=0.40$, suggesting a relatively substantial degree of agreement among the physicians in diagnosing endometriosis (Tables 2–4).

We repeated the interrater reliability analyses after excluding women with poor digital image quality determined a priori by Utah ENDO Study's principal investigators (n=17), and observed similar levels of agreement for endometriosis diagnosis (Fleiss's $\kappa=0.71$, 95% CI 0.66–0.76), rater-based staging of endometriosis (Fleiss's $\kappa=0.45$, 95% CI 0.41–0.48), and computer-assisted staging of endometriosis (Fleiss's $\kappa=0.45$, 95% CI 0.41–0.48). Similar levels of agreement were found after excluding women with poor digital image quality as determined by the eight expert raters (n=50) for endometriosis diagnosis (Fleiss's $\kappa=0.69$, 95% CI 0.64–0.74) and rater-based staging of endometriosis (Fleiss's $\kappa=0.42$, 95% CI 0.39–0.45); however computer-assisted staging of endometriosis dropped from moderate to fair (Fleiss's $\kappa=0.27$, 95% CI 0.22–0.32).

Intra-rater agreement for staging based upon clinical expert assessment versus computer-assisted staging was high among all raters with an average of approximately 90% of women identically classified (range: 83%–97%) and almost perfect agreement (mean weighted $\kappa=0.95$, range: 0.89–0.99). The κ ranges were similar irrespective of type of rater, i.e., weighted κ ranges 0.92 to 0.98 for academic experts and 0.89 to 0.99 for local, specialized expert surgeons).

The agreement between the raters' and computer-assisted staging as depicted in Bland-Altman plots¹⁹ is shown in Figures 1 and 2 for local and expert surgeons, respectively. There was little indication of bias between raters as evidenced by the exclusion of 0 in the 95% CI for the majority of differences for endometriosis total scores. Among all pairwise plots, divergence in differences for endometriosis total scores was found with increasing revised ASRM scores, noticeably after a score of 20. Limits of agreement between any two raters ranged from 38 points below to 49 points above, with an average of 32 points below to 33 points above.

DISCUSSION

We found substantial interrater reliability across raters for endometriosis diagnosis, moderate interrater reliability for endometriosis staging, and almost perfect intrarater reliability for surgeon's experienced assessment versus computer-assisted revised ASRM staging. Reliability of endometriosis diagnosis was 21% higher for academic expert

compared to local, specialized expert surgeons. Combined, our findings support the reliability of endometriosis diagnosis and, to a lesser extent, severity of disease as determined by gynecological surgeons, particularly those with extensive experience.

Our interrater reliability of endometriosis diagnosis agrees with the one previous study conducted on a heterogeneous sample of women. Similar to our study, Weijenborg et al. found substantial interrater reliability among nine senior gynecologists practicing at one medical center for endometrial diagnosis after reviewing 83 videotaped laparoscopies ($\kappa=0.75$, 95% CI 0.59–0.89).⁷ While their study lacked clinical diversity,⁷ we purposely included surgeons from multiple centers with varying levels of experience. Our findings suggest that reliability is not altered substantially by location or composition of clinicians, supporting the conduct of multi-site studies and compilation of endometriosis data across clinical centers.

Hornstein et al. assessed the interrater reliability of endometriosis staging among five sub-specialty reproductive endocrinologists reviewing 20 women with an endometriosis diagnosis using the rAFS⁴ scoring system.⁸ While the surgeons classified the majority (60%) of women at the same stage, the mean interrater agreement was fair ($\kappa=0.28$, range 0.00 to 0.40) compared to the moderate agreement found by Weijenborg et al. using the rAFS system ($\kappa=0.59$, 95% CI 0.43–0.75)⁷ and in our study using the revised ASRM criteria. The small number of assessments per rater ($n=20$) in Hornstein's study, along with a study population restricted to women previously diagnosed with endometriosis, makes comparability between studies difficult.

Our evaluation of the intrarater agreement for the staging of endometriosis by expert assessment versus computer-assisted staging is novel, and intended to remove errors in the application of the revised ASRM criteria on the part of clinicians. Staging agreed for 90% of women with endometriosis and suggests the ability of experienced gynecologists to assess endometriosis stage intuitively without enumeration of the revised ASRM scoring criteria. However, as raters were asked to review revised ASRM criteria prior to reviewing images and knew that their work would be empirically analyzed, the extent to which our findings apply to less experienced surgeons or in other locations remains to be established.

While promising, our findings underscore the need to improve reliability, particularly in regards to the staging of disease irrespective of clinical expertise. We recognize that endometriosis classification systems along with prognostication in regards to pain, fertility, and associated symptoms are in an active state of transformation.^{20–24} For example, a novel pathophysiological-based, functionally-focused classification system in development is the endometriosis fertility index for predicting pregnancy.^{21,24} The optimal system will likely include biomarkers, novel imaging modalities, assessment of anatomic functionality, and outcome measurements in order to accurately prognosticate pain, quality of life, long-term outcomes, as well as the effects of new interventions. The information provided by this study serves as a basis for comparison as new scoring systems are developed.

Our study had several major strengths including the use of a population-based, heterogeneous sample of women for review and an adequate number of blinded assessments

among physicians with varying degrees of experience. Nevertheless, the study had several limitations including variable digital format and quality of surgical visualization documentation. While not ideal, by design, we did not interfere with clinical practice for this observational cohort but rather captured endometriosis as it was being currently diagnosed. Additionally, while we purposely chose reviewers who were academic expert surgeons with fellowship training and current practices in a variety of North American centers, our local, specialized expert surgeons were restricted to a relatively small geographical area, limiting generalizability to other less experienced surgeons practicing at other localities. Finally, because the revised ASRM scoring system does not allow a detailed analysis of the extent of invasive disease that is only accomplished after dissection, our study did not attempt to study newer systems for the staging of invasive disease.

In summary, the reliability of endometriosis diagnosis was substantial and moderate for staging of disease. The slightly higher agreement of computer-assisted versus intuitive staging suggests that future studies on endometriosis may choose to employ checklists to improve reliability between reviewers. These findings may provide reassurance for gynecologists who depend upon images of apparent endometriosis from women receiving care from a variety of practitioners and researchers may infer that studies incorporating multiple sites reviewed by expert surgeons should have a good degree of agreement. It remains to be demonstrated whether additional clinical data (operative reports, histopathology, or primary surgeon interpretation) improves reliability. The data obtained in this study, using current metrics, suggests that endometriosis diagnosis is reliable and staging has room to improve. How the staging of disease burden correlates with multiple clinical outcomes, however, remains to be developed. The ability to maintain reliability in the diagnosis and improve the staging of endometriosis using origin and natural history, as well as functional outcomes, will be critical to meaningful clinical research assessments that will result in future improved outcomes.

Acknowledgements:

The authors thank Denise Lamb, research coordinator, who helped collate and develop the online review system, Lorin Hardy and Sylvia Jessen from the University of Utah School of Medicine Computer Support for designing/developing the online review system, and the NICHD ENDO Study Working Group database managers Christina Bryant and Jansen Davis for their help in extracting and tabulating the information from the online review system.

Funded by the Intramural Research Program, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health (contracts NO1-DK-6-3428; NO1-DK-6-3427; 10001406-02). Ethicon Endo-Surgery, LLC, generously donated the HARMONIC ACE 36P shears and scalpel blades for use in the study through a signed Materials Transfer Agreement with the University of Utah and the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development.

REFERENCES

1. Kennedy S, Bergqvist A, Chapron C, et al. ESHRE guideline for the diagnosis and treatment of endometriosis. *Hum Reprod* 2005;20:2698–2704. [PubMed: 15980014]
2. Practice Committee of the American Society for Reproductive Medicine. Endometriosis and infertility. *Fertil Steril* 2006;86:S156–60. [PubMed: 17055813]
3. Corson SL, Batzer FR, Gocial B, Kelly M, Gutmann JN, Maislin G. Intra-observer and inter-observer variability in scoring laparoscopic diagnosis of pelvic adhesions. *Hum Reprod* 1995;10:161–4. [PubMed: 7745048]

4. Classification of endometriosis. The American Fertility Society. *Fertil Steril* 1979;32:633–4. [PubMed: 510564]
5. Revised American Fertility Society classification of endometriosis: 1985. *Fertil Steril* 1985;43:351–2. [PubMed: 3979573]
6. Revised American Society for Reproductive Medicine classification of endometriosis: 1996. *Fertil Steril* 1997;67:817–821. [PubMed: 9130884]
7. Weijnenborg PT, ter Kuile MM, Jansen FW. Intraobserver and interobserver reliability of videotaped laparoscopy evaluations for endometriosis and adhesions. *Fertil Steril* 2007;87:373–380. [PubMed: 17141769]
8. Hornstein MD, Gleason RE, Orav J, et al. The reproducibility of the revised American Fertility Society classification of endometriosis. *Fertil Steril* 1993;59:1015–1021. [PubMed: 8486168]
9. Rock JA. The revised American Fertility Society classification of endometriosis: reproducibility of scoring. ZOLADEX Endometriosis Study Group. *Fertil Steril* 1995;63:1108–1110. [PubMed: 7720925]
10. Buchweitz O, Wulff P, Malik E. Interobserver variability in the diagnosis of minimal and mild endometriosis. *Eur J Obstet Gynecol Reprod Biol* 2005;122:213–17. [PubMed: 16219522]
11. Farquhar CM. Extracts from the “clinical evidence”. *Endometriosis*. *BMJ* 2000;320:1449–1452. [PubMed: 10827052]
12. Buck Louis GM, Hediger ML, Peterson CM, et al. Incidence of endometriosis by study population and diagnostic method: the ENDO study. *Fertil Steril* 2011;96:360–5. [PubMed: 21719000]
13. Somigliana E, Vercellini P, Vigano P, Benaglia L, Crosignani PG, Fedele L. Non-invasive diagnosis of endometriosis: the goal or own goal? *Hum Reprod* 2010;25:1863–8. [PubMed: 20519246]
14. Ball E, Koh C, Janik G, Davis C. Gynaecological laparoscopy: ‘see and treat’ should be the gold standard. *Curr Opin Obstet Gynecol* 2008;20:325–30. [PubMed: 18660682]
15. Fraser IS. Recognising, understanding and managing endometriosis. *J Hum Reprod Sci* 2008;1:56–64. [PubMed: 19562047]
16. Fleiss JL, Levin BA, Paik MC. *Statistical methods for rates and proportions*. 3rd ed. Hoboken (NJ): Wiley-Interscience; 2003.
17. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37–46.
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–174. [PubMed: 843571]
19. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Int J Nurs Stud* 2010;47:931–6.
20. Roberts CP, Rock JA. The current staging system for endometriosis: does it help? *Obstet Gynecol Clin North Am* 2003;30:115–132. [PubMed: 12699261]
21. Adamson GD, Pasta DJ. Endometriosis fertility index: the new, validated endometriosis staging system. *Fertil Steril*. 2010;94:1609–1615.
22. Haas D, Chvatal R, Habelsberger A, Wurm P, Schimetta W, Oppelt P. Comparison of revised American Fertility Society and ENZIAN staging: a critical evaluation of classifications of endometriosis on the basis of our patient population. *Fertil Steril* 2011;95:1574–78. [PubMed: 21315335]
23. Coccia ME, Rizzello F. Ultrasonographic staging: a new staging system for deep endometriosis. *Ann N Y Acad Sci* 2011;1221:61–9. [PubMed: 21401631]
24. Adamson GD. Endometriosis classification: an update. *Curr Opin Obstet Gynecol* 2011;23:213–20. [PubMed: 21666464]

Figure 1:

Agreement between the local, specialized expert surgeons regarding the severity and extent of endometriosis, as assessed via the computer-assisted revised American Society for Reproductive Medicine (ASRM) score. Mean difference and 95% confidence interval (CI) reported for each pair-wise comparison.

Figure 2:

Agreement between the academic expert surgeons regarding the severity and extent of endometriosis, as assessed via the computer-assisted revised American Society for Reproductive Medicine (ASRM) score. Mean difference and 95% confidence interval (CI) reported for each pair-wise comparison.

Table 1:

Comparison of Study Participants by Endometriosis Diagnosis

Characteristic	Endometriosis Diagnosis n=105	No Endometriosis Diagnosis n=43	<i>P</i> [*]
Age (mean ± SD)	31.4 ± 6.5	33.3 ± 7.1	0.17
Race (n [%])			
Hispanic	13 (12.4)	3 (7.0)	0.46
Non-Hispanic white	83 (79.0)	36 (83.7)	
Hispanic black	1 (1.0)	1 (2.3)	
Asian, Islander, or Native	4 (3.8)	3 (7.0)	
Other or multiracial	4 (3.8)	0 (0.0)	
Married or living as married (n [%])	85 (81.7)	38 (88.4)	0.32
Household income			
Below poverty line	7 (6.9)	7 (16.3)	0.21
Within 180% of poverty	5 (4.9)	2 (4.7)	
Above poverty	90 (88.2)	34 (79.1)	
Previous pregnancy (n [%])	54 (51.4)	34 (79.1)	0.002
Previous live birth (n [%])	40 (38.1)	32 (74.4)	<0.001
Primary reason for surgery			
Tubal ligation	5 (4.8)	4 (9.3)	0.23
Pelvic pain	70 (66.7)	22 (51.2)	
Pelvic mass	15 (14.3)	4 (9.3)	
Infertility	6 (5.7)	5 (11.6)	
Leiomyomas	2 (1.9)	2 (4.7)	
Menstrual irregularities	7 (6.7)	6 (14.0)	

* Categorical variables with chi-square test, continuous variable with nonparametric Wilcoxon test.

SD, standard deviation.

Table 2:

Raters' (n=8) * Interrater Agreement for the Presence or Absence of Endometriosis After Evaluating Digital Images (Percentage Agreement, Kappa Statistic and Confidence Interval for Kappa)

% Kappa κ (95% CI)	1							
1	1.0	2						
2	81% 0.60 (0.45,0.75) [†]	1.0	3					
3	79% 0.57 (0.43,0.72) [†]	84% 0.61 (0.46,0.77) [†]	1.0	4				
4	76% 0.47 (0.32,0.61)	93% 0.81 (0.67,0.94) [†]	79% 0.50 (0.35,0.65)	1.0	5			
5	80% 0.60 (0.44,0.75) [†]	89% 0.73 (0.59,0.88) [†]	81% 0.58 (0.43,0.74) [†]	87% 0.68 (0.53,0.83) [†]	1.0	6		
6	80% 0.59 (0.45,0.73) [†]	86% 0.68 (0.54,0.83) [†]	82% 0.62 (0.50,0.75) [†]	84% 0.60 (0.45,0.74) [†]	91% 0.81 (0.69,0.92) [†]	1.0	7	
7	88% 0.74 (0.59,0.88) [†]	87% 0.69 (0.52,0.85) [†]	79% 0.54 (0.38,0.70) [†]	87% 0.65 (0.47,0.82) [†]	88% 0.71 (0.54,0.88) [†]	89% 0.76 (0.64,0.89) [†]	1.0	8
8	83% 0.65 (0.49,0.79) [†]	89% 0.72 (0.58,0.87) [†]	83% 0.61 (0.46,0.76) [†]	89% 0.69 (0.55,0.84) [†]	94% 0.86 (0.76,0.97) [†]	91% 0.79 (0.67,0.91) [†]	95% 0.86 (0.74,0.98) [†]	1.0

* Raters 1–4 are local specialized surgeons; 5–8 are academic expert surgeons.

[†] $P < 0.05$ (Test of $H_0: \kappa = 0.4$ against $H_1: \kappa = 0.75$).

Table 3:

Raters' (n=8) * Interrater Agreement for Stage of Endometriosis, Using Assessment Based on Expert Scoring After Evaluating Digital Images (Percentage Agreement, Kappa Statistic and Confidence Interval for Kappa).

†

% Kappa κ (95% CI)	1							
1	1.0	2						
2	55% 0.62 (0.46,0.77)	1.0	3					
3	54% 0.58 (0.42,0.74)	48% 0.60 (0.45,0.75)	1.0	4				
4	51% 0.72 (0.61,0.83)	67% 0.78 (0.66,0.91)	50% 0.59 (0.45,0.73)	1.0	5			
5	62% 0.64 (0.47,0.81)	57% 0.70 (0.59,0.82)	62% 0.68 (0.53,0.83)	64% 0.76 (0.66,0.86)	1.0	6		
6	59% 0.66 (0.53,0.78)	49% 0.64 (0.51,0.76)	60% 0.66 (0.55,0.78)	51% 0.66 (0.55,0.78)	68% 0.72 (0.60,0.83)	1.0	7	
7	60% 0.68 (0.52,0.84)	55% 0.72 (0.58,0.85)	55% 0.61 (0.46,0.77)	54% 0.71 (0.58,0.83)	60% 0.75 (0.62,0.87)	58% 0.70 (0.59,0.81)	1.0	8
8	60% 0.68 (0.55,0.81)	62% 0.71 (0.57,0.85)	57% 0.63 (0.46,0.79)	65% 0.78 (0.68,0.87)	69% 0.83 (0.75,0.90)	59% 0.69 (0.57,0.81)	60% 0.77 (0.67,0.87)	1.0

* Raters 1–4 are local specialized surgeons; 5–8 are academic expert surgeons.

† All significant at $P < 0.05$ (Test of $H_0: \kappa = 0.4$ against $H_1: \kappa = 0.75$)

Table 4:

Raters' (n=8) * Interrater Agreement for Stage of Endometriosis, Using Assessment Based on Computer-Assisted American Society for Reproductive Medicine Scoring After Evaluating Digital Images (Percentage Agreement, Kappa Statistic and Confidence Interval for Kappa).

% Kappa κ (95% CI)	1							
1	1.0	2						
2	61% 0.64 (0.49,0.80) [†]	1.0	3					
3	61% 0.56 (0.38,0.74)	63% 0.61 (0.44,0.78) [†]	1.0	4				
4	55% 0.71 (0.59,0.82) [†]	71% 0.77 (0.64,0.72) [†]	54% 0.58 (0.43,0.73) [†]	1.0	5			
5	61% 0.60 (0.42,0.79) [†]	62% 0.70 (0.57,0.83) [†]	70% 0.71 (0.54,0.88) [†]	62% 0.76 (0.65,0.86) [†]	1.0	6		
6	58% 0.62 (0.48,0.77) [†]	66% 0.66 (0.52,0.79) [†]	68% 0.69 (0.58,0.81) [†]	52% 0.64 (0.51,0.77) [†]	75% 0.77 (0.66,0.88)	1.0	7	
7	63% 0.69 (0.53,0.85) [†]	57% 0.78 (0.65,0.91) [†]	59% 0.66 (0.49,0.82) [†]	54% 0.70 (0.57,0.83) [†]	63% 0.76 (0.63,0.89)	55% 0.68 (0.55,0.80)	1.0	8
8	61% 0.65 (0.52,0.77) [†]	61% 0.65 (0.50,0.80) [†]	58% 0.55 (0.37,0.73)	64% 0.73 (0.64,0.83) [†]	71% 0.80 (0.72,0.88)	57% 0.66 (0.54,0.79)	62% 0.69 (0.57,0.82)	1.0

* Raters 1–4 are local specialized surgeons; 5–8 are academic expert surgeons.

[†] P<0.05 (Test of H₀: κ = 0.4 against H₁: κ = 0.75)