# Interrater and Intrarater Reliability in the Diagnosis and Staging of Endometriosis

*Karen C. Schliep, PhD, MSPH, Joseph B. Stanford, MD, MSPH, Zhen Chen, PhD, Bo Zhang, PhD, Jessie K. Dorais, MD, Erica Boiman Johnstone, MD, Ahmad O. Hammoud, MD, Michael W. Varner, MD, Germaine M. Buck Louis, PhD, and C. Matthew Peterson, MD, on behalf of the Endometriosis: Natural History, Diagnosis and Outcomes (ENDO) Study Working Group*

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

**METHODS:** The 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-level 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% confidence interval [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).

**CONCLUSION:** Substantial reliability was found for revised ASRM endometriosis diagnosis, whereas moderate reliability was observed for staging. Almost perfect reliability was observed for surgeons' rating of disease severity compared with 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. Although surgeons appear to be skilled at assessing endometriosis stage intuitively, how staging of disease burden correlates with clinical outcomes remains to be developed.

*(Obstet Gynecol 2012;120:104–12)*
*DOI: 10.1097/AOG.0b013e31825bc6cf*

**LEVEL OF EVIDENCE:** II

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 misclas-

*From the Department of Family and Preventive Medicine and the Divisions of Reproductive Endocrinology and Infertility and Maternal-Fetal Medicine, Department of Obstetrics and Gynecology, and the Division of Maternal-Fetal Medicine, University of Utah, Salt Lake City, Utah; the Division of Epidemiology, Statistics and Prevention Research,* Eunice Kennedy Shriver *National Institute of Child Health and Human Development, the National Institutes of Health, Rockville, Maryland; and School of Biological and Population Health Sciences, College of Public Health and Human Sciences, Oregon State University, Corvallis, Oregon.*

*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.*

*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 and developing the online review system; and the* Eunice Kennedy Shriver *National Institute of Child Health and Human Development 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.*

*Corresponding author: C. Matthew Peterson, MD, Division of Reproductive Endocrinology and Infertility, Department of Obstetrics and Gynecology, 30 North 1900 East, Suite 2B200, University of Utah, Salt Lake City, UT 84132; e-mail: c.matthew.peterson@hsc.utah.edu.*

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


ISSN: 0029-7844/12

sification bias of endometriosis status among gynecologists threatens our ability to better understand the etiology and clinical management of the disease.

Although 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 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 hospital[7–9] or nonexperts.[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 multisite 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 study.[12]

## MATERIALS AND METHODS

A random sample of women was selected for study from the larger Endometriosis: Natural History, Diagnosis and Outcomes study for the specific aim of conducting a reliability study. Briefly, the study used a matched exposure cohort design to assess environmental chemicals and lifestyle behaviors associated with endometriosis in two study cohorts: operative and population.[12] Endometriosis: Natural History, Diagnosis and Outcomes 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 on the operative cohort in which 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 or 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 photographs 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. Because there was a spectrum of image quality, Endometriosis: Natural History, Diagnosis and Outcomes 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 Endometriosis: Natural History, Diagnosis and Outcomes principal investigators (C.M.P., J.B.S.).

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 greater than 99% and greater than 85% statistical power for testing interrater reliability for the presence or absence of endometriosis and staging (between I-II and III-IV), respectively. Power calculations were based on 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, whereas 21% of women without an endometriosis diagnosis had good-quality images.

Four academic expert surgeons and four local, specialized expert (ie, fellowship-trained) surgeons, determined a priori by Endometriosis: Natural History, Diagnosis and Outcomes study principal investigators, were recruited for the study. As a result of 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.

Endometriosis: Natural History, Diagnosis and Outcomes 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 through 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. Before 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; and 3) the computer-assisted staging based on the expert's checklist of findings and the revised ASRM algorithm.[6]

Descriptive statistics were calculated to summarize rater characteristics by rater type, expert compared with local gynecologic surgeons. Interrater agreement for the clinical diagnosis of endometriosis was evaluated by $\kappa$ statistics.[16,17] The $\kappa$ 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, $\kappa=1$. When $\kappa$ equals 0, the agreement is no better than what would be obtained by chance alone. For our reliability analyses, we used Landis and Koch's[18] guidelines for interpreting $\kappa$ statistics: $\kappa$ between 0.00 and 0.20 indicated slight agreement; $\kappa$ between 0.21 and 0.40 denoted fair agreement; $\kappa$ between 0.41 and 0.60 characterized moderate agreement; $\kappa$ between 0.61 and 0.80 defined substantial agreementl and a value of $\kappa$ greater than 0.80 equated to almost perfect agreement. For binary outcomes (eg, presence or absence of endometriosis), we computed pairwise agreement using Cohen's $\kappa$ and multirater agreement using Fleiss' multirater $\kappa$. For ordinal outcomes (eg, staging of endometriosis), we computed pairwise agreement using Cohen's weighted $\kappa$ with squared weights and multirater agreement using Fleiss' multirater $\kappa$. For continuous outcome (ie, revised ASRM score), we constructed pairwise Bland-Altman plots[19] 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 $\kappa$ and their 95% confidence intervals (CIs) were estimated. For each $\kappa$, the $P$ value was calculated from the one-sided Wald test with null hypothesis $\kappa=0.40$ compared with alternative hypothesis $\kappa=0.75$. If a woman had an "indeterminate" diagnosis for endometriosis, that woman was excluded from the final data analysis under the assumption of data being missing at random. Analyses were performed in R 2.13.1 and SAS 9.2.

## RESULTS

Eight raters (four academic expert and four local, specialized expert gynecologic surgeons) assessed images during the summer of 2010. Although 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, interquartile range 15.0–21.0, range 15–21 years) compared with the local, specialized expert surgeons (median 5.0, interquartile range 3.5–12.5, range 2–20 years). Academic expert surgeons also had more experience authoring, teaching, and serving as principal or coinvestigator of funded studies addressing endometriosis compared with local, specialized expert surgeons. Neither the number of patients with endometriosis seen per week nor the number of laparoscopies performed per month substantially differed between the rater groups. Although

there were no significant differences between groups in regard to age, race, marital status, income, and primary reason for surgery, women with an endometriosis diagnosis were less likely to previously been pregnant ($P$=.002) or have had a live birth ($P$<.001) compared with women without an endometriosis diagnosis (Table 1). Among the 148 women, 121 (82%) had only digital photographs, three (2%) had only digital video images, and 24 (16%) had both digital photographs 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, ie, 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).

**Table 1. Comparison of Study Participants by Endometriosis Diagnosis**

| Characteristic | Endometriosis Diagnosis (n=105) | No Endometriosis Diagnosis (n=43) | P* |
|---|---|---|---|
| Age (y) | 31.4±6.5 | 33.3±7.1 | .17 |
| Race | | | |
| Hispanic | 13 (12.4) | 3 (7.0) | .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 | 85 (81.7) | 38 (88.4) | .32 |
| Household income | | | |
| Below poverty line | 7 (6.9) | 7 (16.3) | .21 |
| Within 180% of poverty | 5 (4.9) | 2 (4.7) | |
| Above poverty | 90 (88.2) | 34 (79.1) | |
| Previous pregnancy | 54 (51.4) | 34 (79.1) | .002 |
| Previous live birth | 40 (38.1) | 32 (74.4) | <.001 |
| Primary reason for surgery | | | |
| Tubal ligation | 5 (4.8) | 4 (9.3) | .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) | |

Data are mean±standard deviation or n (%) unless otherwise specified.

* Categorical variables with $\chi^2$ test; continuous variable with nonparametric Wilcoxon test.

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' multirater $\kappa$ of 0.69 (95% CI 0.64–0.74). The academic expert surgeons had substantial interrater reliability (Fleiss' $\kappa$=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 (ie, 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 compared with 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 the Utah Endometriosis: Natural History, Diagnosis and Outcomes study's principal investigators (n=17) and observed similar levels of agreement for endometriosis diagnosis (Fleiss' $\kappa$=0.71, 95% CI 0.66–0.76), rater-based staging of endometriosis (Fleiss' $\kappa$=0.45, 95% CI 0.41–0.48), and computer-assisted staging of endometriosis (Fleiss' $\kappa$=0.45, 95% CI 0.41–0.48). Similar levels of agreement were found after excluding women with

**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)**

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

CI, confidence interval.
Data are % and kappa (95% CI). Numbers 1–8 in bold represent the raters.
* Raters 1–4 are local specialized surgeons; raters 5–8 are academic expert surgeons.
† $P<.05$ (test of $H_0$: $\kappa=0.4$ against $H_1$: $\kappa=0.75$).

poor digital image quality as determined by the eight expert raters (n=50) for endometriosis diagnosis (Fleiss' $\kappa=0.69$, 95% CI 0.64–0.74) and rater-based staging of endometriosis (Fleiss' $\kappa=0.42$, 95% CI 0.39–0.45); however, computer-assisted staging of endometriosis dropped from moderate to fair (Fleiss' $\kappa=0.27$, 95% CI 0.22–0.32).

Intrarater agreement for staging based on clinical expert assessment compared with 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 $\kappa$ ranges were similar irrespective of type of rater, ie, weighted $\kappa$ ranges from 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 plots[19] 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 inclusion 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

**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)†**

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

CI, confidence interval.
Data are % and kappa (95% CI). Numbers 1–8 in bold represent the raters.
* Raters 1–4 are local specialized surgeons; raters 5–8 are academic expert surgeons.
† All significant at $P<.05$ (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)**

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

CI, confidence interval.
Data are % and kappa (95% CI). Numbers 1–8 in bold represent the raters.
* Raters 1–4 are local specialized surgeons; raters 5–8 are academic expert surgeons.
† $P<.05$ (test of $H_0$: $\kappa=0.4$ against $H_1$: $\kappa=0.75$).

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 compared with computer-assisted revised ASRM staging. Reliability of endometriosis diagnosis was 21% higher for academic expert compared with local, specialized expert surgeons. Combined, our findings support the reliability of endometriosis diagnosis and, to a lesser extent, severity of disease as determined by gynecologic 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[7] 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). Although 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 multisite studies and compilation of endometriosis data across clinical centers.

Hornstein et al[8] assessed the interrater reliability of endometriosis staging among five subspecialty reproductive endocrinologists reviewing 20 women with an endometriosis diagnosis using the revised American Fertility Society[4] scoring system. Although the surgeons classified the majority (60%) of women at the same stage, the mean interrater agreement was fair ($\kappa=0.28$, range 0.00–0.40) compared with the moderate agreement found by Weijenborg et al[7] using the revised American Fertility Society system ($\kappa=0.59$, 95% CI 0.43–0.75) and in our study using the revised ASRM criteria. The small number of assessments per rater (n=20) in Hornstein et al'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 compared with 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, because raters were asked to review revised ASRM criteria before 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.

Although promising, our findings underscore the need to improve reliability, particularly in regard to the staging of disease irrespective of clinical expertise. We recognize that endometriosis classification systems along with prognostication in regard to pain,

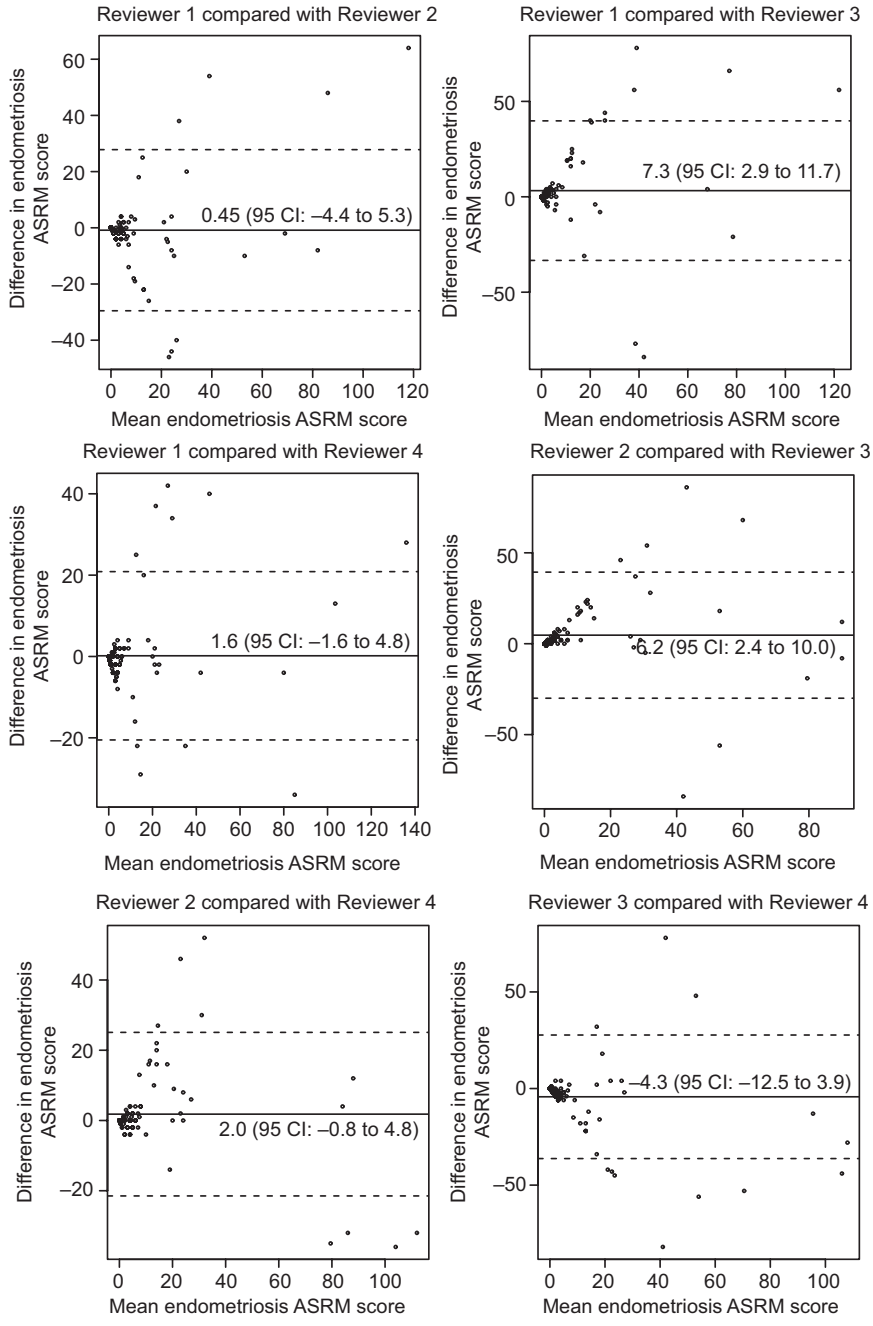


**Fig. 1.** Agreement between the local, specialized expert surgeons regarding the severity and extent of endometriosis as assessed through the computer-assisted revised American Society for Reproductive Medicine (ASRM) score. Mean difference and 95% confidence interval (CI) reported for each pairwise comparison.

*Schliep. Endometriosis Diagnosis and Staging Reliability. Obstet Gynecol 2012.*

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 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 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. Although not ideal, by design, we did not interfere

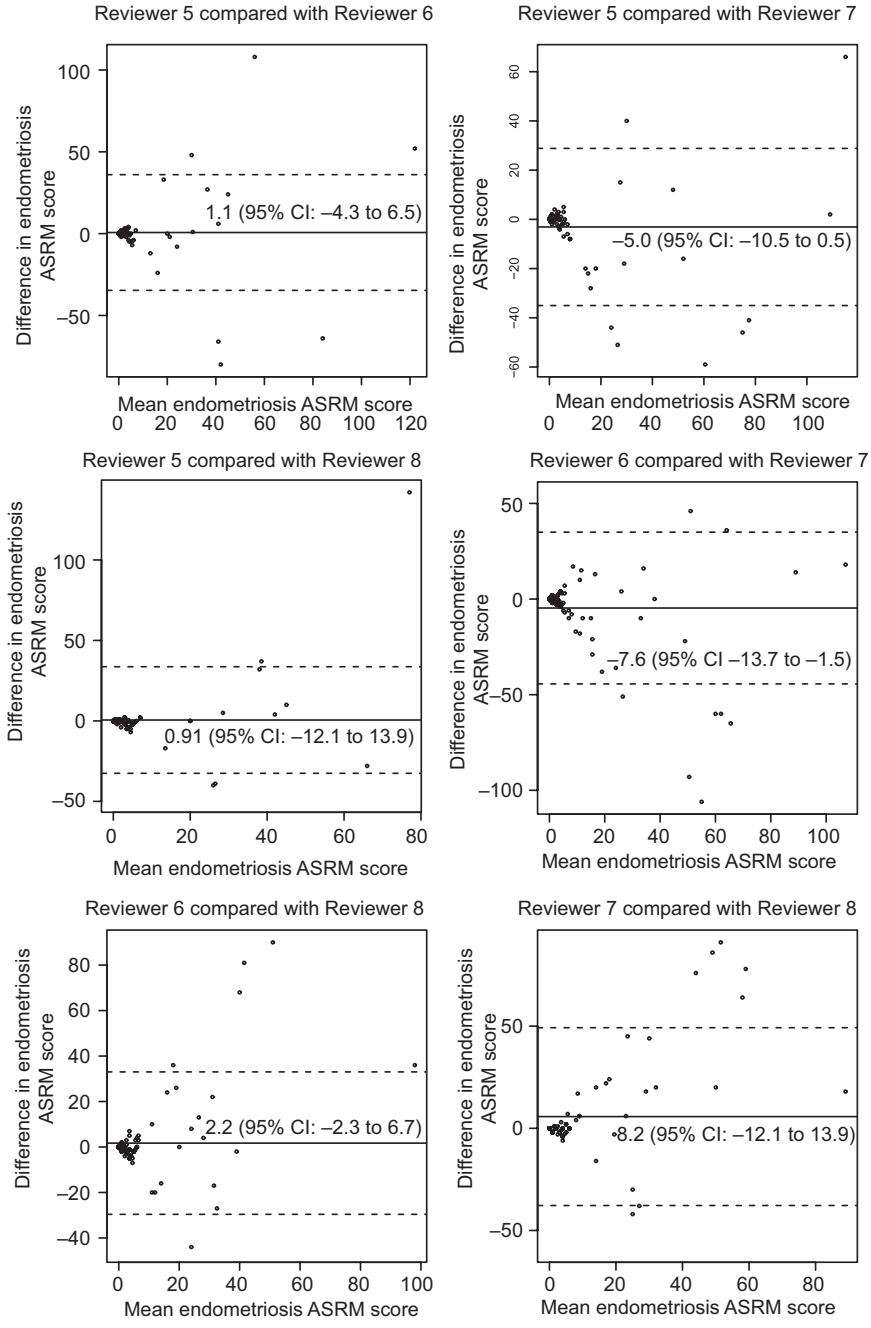


**Fig. 2.** Agreement between the academic expert surgeons regarding the severity and extent of endometriosis as assessed through the computer-assisted revised American Society for Reproductive Medicine (ASRM) score. Mean difference and 95% confidence interval (CI) reported for each pairwise comparison.

*Schliep. Endometriosis Diagnosis and Staging Reliability. Obstet Gynecol 2012.*

with clinical practice for this observational cohort but rather captured endometriosis as it was being currently diagnosed. Additionally, although 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 compared with intuitive staging suggests that future studies on endometriosis may choose to use

checklists to improve reliability between reviewers. These findings may provide reassurance for gynecologists who depend on images of apparent endometriosis from women receiving care from a variety of health care 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, suggest 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.

## REFERENCES


1. Kennedy S, Bergqvist A, Chapron C, D'Hooghe T, Dunselman G, Greb R, et al. ESHRE guideline for the diagnosis and treatment of endometriosis. Hum Reprod 2005;20:2698–704.
2. Practice Committee of the American Society for Reproductive Medicine. Endometriosis and infertility. Fertil Steril 2006;85: S156–60.
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.
4. Classification of endometriosis. The American Fertility Society. Fertil Steril 1979;32:633–4.
5. Revised American Fertility Society classification of endometriosis: 1985. Fertil Steril 1985;43:351–2.
6. Revised American Society for Reproductive Medicine classification of endometriosis: 1996. Fertil Steril 1997;67:817–21.
7. Weijenborg PT, ter Kuile MM, Jansen FW. Intraobserver and interobserver reliability of videotaped laparoscopy evaluations for endometriosis and adhesions. Fertil Steril 2007;87:373–80.
8. Hornstein MD, Gleason RE, Orav J, Haas ST, Friedman AJ, Rein MS, et al. The reproducibility of the revised American Fertility Society classification of endometriosis. Fertil Steril 1993;59:1015–21.
9. Rock JA. The revised American Fertility Society classification of endometriosis: reproducibility of scoring. ZOLADEX Endometriosis Study Group. Fertil Steril 1995;63:1108–10.
10. Buchweitz O, Wulfing P, Malik E. Interobserver variability in the diagnosis of minimal and mild endometriosis. Eur J Obstet Gynecol Reprod Biol 2005;122:213–7.
11. Farquhar CM. Extracts from the 'clinical evidence.' Endometriosis. BMJ 2000;320:1449–52.
12. Buck Louis GM, Hediger ML, Peterson CM, Croughan M, Sundaram R, Stanford J, et al. Incidence of endometriosis by study population and diagnostic method: the ENDO study. Fertil Steril 2011;96:360–5.
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.
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.
15. Fraser IS. Recognising, understanding and managing endometriosis. J Hum Reprod Sci 2008;1:56–64.
16. Fleiss JL, Levin BA, Paik MC. Statistical methods for rates and proportions. 3rd ed. Hoboken (NJ): Wiley-Interscience; 2003.
17. Cohen JA. 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;33:159–74.
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: 115–32.
21. Adamson GD, Pasta DJ. Endometriosis fertility index: the new, validated endometriosis staging system. Fertil Steril 2010;94: 1609–15.
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;95:1574–8.
23. Coccia ME, Rizzello F. Ultrasonographic staging: a new staging system for deep endometriosis. Ann N Y Acad Sci 2011; 1221:61–9.
24. Adamson GD. Endometriosis classification: an update. Curr Opin Obstet Gynecol 2011;23:213–20.