Published in final edited form as:

BJOG. 2017 January ; 124(2): 220–229. doi:10.1111/1471-0528.13711.

Endometriosis diagnosis and staging by operating surgeon and expert review using multiple diagnostic tools: an interrater agreement study

Karen C. Schliep, PhD[a,b], **Zhen Chen, PhD**[a], **Joseph B. Stanford, MD**[b], **Yunlong Xie, PhD**[a], **Sunni L. Mumford, PhD**[a], **Ahmad O. Hammoud, MD**[c], **Erica Boiman Johnstone, MD**[c], **Jessie K. Dorais, MD**[c], **Michael W. Varner, MD**[c], **Germaine M. Buck Louis, PhD**[a], and **C. Matthew Peterson, MD**[c]

[a]Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Rockville, Maryland
[b]Department of Family and Preventive Medicine, University of Utah, Salt Lake City, Utah
[c]Department of Obstetrics and Gynecology, University of Utah, Salt Lake City, Utah

Abstract

Objective—To determine agreement of endometriosis diagnosis between real-time laparoscopy and subsequent expert review of digital images, operative reports, magnetic resonance imaging (MRI) and histopathology, viewed sequentially.

Design—Interrater agreement study

Setting—Five urban surgical centers

Population—Women, aged 18–44 years, who underwent a laparoscopy regardless of clinical indication. A random sample of 105 women with and 43 without a postoperative endometriosis diagnosis was obtained from the ENDO Study.

Methods—Laparoscopies were diagnosed, digitally recorded, and reassessed.

Correspondence: Karen Schliep, Division of Public Health, Department of Family and Preventive Medicine, University of Utah, 375 Chipeta Way, Suite A, Salt Lake City, UT 84108. Phone: 801-587-3315; karen.schliep@utah.edu.

CONTRIBUTION TO AUTHORSHIP

This add-on study to the main ENDO Study required input from multiple collaborators, all of whom met criteria for authorship. Specifically, Drs. Schliep and Chen had full access to all of the data in the study and were responsible for the integrity of the data and the accuracy of the data analysis; Drs. Buck Louis, Chen, Peterson, Stanford, Varner, Boiman Johnstone, Hammoud, and Dorais contributed to the study concept and design and oversaw acquisition of the data and study supervision; Drs. Schliep, Chen, and Xie conducted the statistical analyses, including independent coding and replication of results; and all the authors listed helped with the interpretation of the data and the drafting and editing of the manuscript.

DISCLOSURE OF INTERESTS

None of the authors have a conflict of interest (the ICMJE disclosure forms are available online).

DETAILS OF ETHICAL APPROVAL

Full human subjects approval (Institutional Review Board, University of Utah; Intermountain Healthcare Office of Research, Utah; and the National Institutes of Health Institutional Review Board Reliance) was obtained for the conduct of this study and each of the women provided informed consent before any data collection. Data were deidentified and encrypted before being uploaded to the web-based data system. The University of Utah Institutional Review Board approved this ENDO follow-on study and all physicians signed an informed consent before being given access to the de-identified online review system.

Main Outcome Measures—Interobserver agreement of endometriosis diagnosis and staging according to the revised American Society for Reproductive Medicine criteria. Prevalence and bias adjusted kappas (κ) were calculated for diagnosis and weighted κ were calculated for staging.

Results—Surgeons and expert reviewers had substantial agreement on diagnosis and staging after viewing digital images (n=148) (mean κ=0.67, range: 0.61–0.69; mean κ=0.64, range: 0.53–0.78, respectively) and after additionally viewing operative reports (n=148) (mean κ=0.88, range: 0.85–0.89; mean κ=0.85, range: 0.84–0.86, respectively). While additionally viewing MRI findings (n=36) did not greatly impact agreement, agreement substantially decreased after viewing histological findings (n=67) with expert reviewers changing their assessment from a positive to negative diagnosis in up to 20% of cases.

Conclusions—While these findings suggest that misclassification bias in diagnosis or staging of endometriosis via visualized disease is minimal, they should alert gynecologists who review operative images in order to make decisions on endometriosis treatment that operative reports/drawings and histopathology, but not necessarily MRI, will improve their ability to make sound judgments.

Tweetable abstract—Endometriosis diagnosis and staging agreement between expert reviewers and operating surgeons was substantial.

Keywords

endometriosis; epidemiology; reliability; laparoscopy; histology; magnetic resonance imaging

INTRODUCTION

Endometriosis is a common gynecological disorder affecting at least 11% of reproductive-age women,[1,2] but increasing to approximately half of women experiencing pelvic pain or infertility.[3–5] The true incidence and prevalence at the population level remains unknown, due in part to varying clinical diagnostic proficiency.[6] Endometriosis is difficult to diagnose and prone to misclassification given its differing symptomatology, paucity of consistent physical examination findings, unpredictable disease course, and lack of an identifying biomarker.[1,6,7] Despite the challenges, improving the diagnostic and staging accuracy of endometriosis in women with pelvic pain and infertility is paramount for effectively treating the resulting, sometimes physically or psychologically debilitating, conditions.[6] Furthermore, proper evaluation of potential risk factors or investigational treatments is dependent on consistent diagnosis across clinical centers.

Operative real-time laparoscopic findings using standardized staging systems are considered the current gold standard for diagnosing endometriosis and assessing its severity.[8,9] As per recent guidelines,[10] histopathologic evaluation is recommended for diagnostic confirmation, but its true value has not been adequately quantified due to nonstandardized and unblinded assessment biasing previous studies.[11] Current advice notes that while positive histology can confirm the diagnosis of endometriosis, negative histology does not exclude it.[10]

Recording laparoscopic surgeries via digital imaging has become widely accepted among gynecological surgeons for clinical, research, and medicolegal purposes.[12] However, few studies have assessed the operating surgeon's findings with those of expert reviewers, particularly among a heterogeneous study population for whom laparoscopies were conducted by a diverse group of surgeons practicing at a variety of clinical centers.[12-14] Additionally, while diagnostic accuracy in general is known to improve with added clinical information, no study to date has assessed specifically whether additional information obtained from preoperative magnetic resonance imaging (MRI) and/or postoperative histopathologic examination alters endometriosis diagnosis and staging accuracy.

While we previously reported substantial reliability (κ=0.69) of endometriosis diagnosis between expert consulting gynecologists after viewing digitally recorded laparoscopies from the Endometriosis, Natural History, Diagnosis, and Outcomes (ENDO) Study,[15] the diagnostic agreement for endometriosis between the ENDO Study operating surgeons (n=46) and expert review of digitally recorded laparoscopies augmented by additional clinical information has yet to be clarified. Therefore, our objective for this study was to determine agreement of endometriosis diagnosis and staging between real-time laparoscopy and subsequent expert consultants' review of digital images, operative reports, MRI, and/or histopathologic findings in a random sample of women from the ENDO Study operative cohort. Additionally, we set out to evaluate histopathologic confirmation and MRI consistency of visually diagnosed endometriosis by the operating surgeon.

METHODS

Study Population

Using a block randomization approach, we selected 105 women with and 43 women without a postoperative endometriosis diagnosis, as per gold standard of real-time laparoscopic findings, among the total Utah operative cohort of 473 women who participated in the ENDO Study (2007–2009). Our stratification scheme was developed *a priori* to ensure the study had ≥ 99% power for testing the inter-rater reliability of an endometriosis diagnosis and ≥ 85% power for staging (comparing I-II and III-IV), respectively. Power calculations were based on an α (two-sided) of 0.05 and other assumptions as derived from the literature.

While the goal of the ENDO Study was to assess associations between environmental chemicals, lifestyle behaviors, and endometriosis,[1] the goal of this study, the ENDO Physician Reliability Study, was to assess the inter- and intra-rater agreement among gynecological surgeons with respect to endometriosis diagnosis and staging. The study population, materials, and methods for the ENDO Study have been previously described.[1] Briefly, ENDO Study participants from the Utah operative cohort were recruited from one of five participating hospital surgical centers who were scheduled to undergo a diagnostic and/or therapeutic laparoscopy or laparotomy regardless of clinical indication. To be eligible for study participation, women had to be currently menstruating, aged 18–44 years, with no prior history of surgically diagnosed endometriosis (prevalent cases).

Consistent with the study's observational design, surgeons were not required to change their clinical practice in any way, including decision making about obtaining endometrial implants for histologic review. All women underwent a diagnostic and/or therapeutic laparoscopy, among whom 117 (27%) had endometrial implants sent for blinded histopathological review. Histologically confirmed endometriosis included endometrial glands and/or stroma and/or hemosiderin-laden macrophages. A random sample (n=86; 20%) of women additionally underwent a pelvic MRI prior to surgery. Women were block randomized in a 1:1 ratio for MRI selection based on preoperative diagnosis (suspected versus not suspected endometriosis). One radiologist conducted and read all of the MRIs, using either a Siemens Avanto or Espree 1.5 Tesla scanner and a U.S. Food and Drug Administration-approved protocol for pelvic imaging, and completed standardized data collection instruments. All images were reread by a second radiologist and all endometriosis diagnoses were affirmed. Participants provided informed consent before any data collection and were compensated for their time and travel. Full human subject's approval was obtained from the University of Utah Institutional Review Board and Intermountain Healthcare Office of Research along with a signed IRB Reliance Agreement from the National Institutes of Health.

All 148 women had digital images and operative reports available for analysis. In addition, 36 women (24%) were among those randomized in the ENDO Study to undergo a pelvic MRI (27 diagnosed with endometriosis, 9 diagnosed without endometriosis), 67 women (45%) had endometrial implants sent for histopathologic review (66 diagnosed with endometriosis, 1 diagnosed without endometriosis), and 22 women (15%) underwent both a pelvic MRI and had specimens sent for histopathologic review (22 diagnosed with endometriosis, 1 diagnosed without endometriosis).

Operative Surgeons’ Assessments

Postoperative data collection instruments were obtained from 46 participating gynecologic surgeons, all of whom had surgical training in the diagnosis and staging of endometriosis. As noted above, participating surgeons were not asked to alter clinical practice in arriving at the diagnosis or treatment of endometriosis or other observed gynecologic pathology; however, they were asked to take intraoperative digital photos to document diagnoses. Results of the pelvic MRI examinations done under the ENDO Study protocol were for study purposes and were not routinely available to operating surgeons at the time of the surgery. Surgeons completed a standardized report that captured all operative diagnoses and findings, including endometriosis diagnosis (yes/no). For women diagnosed with endometriosis, surgeons were first asked to empirically stage severity (i.e., experienced assessment without assistance of checklist) via the revised American Society for Reproductive Medicine's (rASRM) criteria (I-IV: minimal, mild, moderate, severe) and then asked to complete the rASRM checklist of staging criteria, from which an algorithm automatically calculated the rASRM weighted point score and stage. The operating surgeons empirically staged 40 (38.1%) women as having minimal, 24 (22.9%) as having mild, 27 (25.7%) as having moderate, and 14 (13.3%) as having severe endometriosis. The distribution of staging for women based upon the automated rASRM algorithm was 57 (54.8%) minimal, 16 (15.4%) mild, 12 (11.5%) moderate, and 19 (18.3%) severe disease.

Expert Reviewers’ Assessments

We recruited four academic expert surgeons from North American clinical centers who had extensive clinical and research experience in the diagnosis and management of endometriosis and who were directors of laparoscopic gynecologic training programs. The University of Utah Institutional Review Board approved this ENDO follow-on study (2010), and all physicians signed an informed consent before being given access to the de-identified online review system.

A trained research nurse prepared anonymous digital images (82% digital photographs, 18% both digital photographs and video) free of all clinical information for the random sample using a standardized format. As in clinical practice, the quality of photographs varied necessitating the need for independent assessment by authors CMP and JBS as good or poor. We intentionally included poor images (n=18) in our random sample to better reflect what clinicians may encounter in clinical practice and to augment generalizability. Specifically, 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.

An online system was designed to mimic the diagnostic information that surgeons would typically have in the consultative review of patients, received sequentially. The expert reviewers first received the digital images and were asked to complete a review based on that information alone. For each woman, raters received the operative report/drawing (indicating operative findings but not including diagnosis or assessment of severity) and were asked to complete the review before receiving the MRI report, and finally, the histology report. At each round of review, the expert reviewer determined the presence or absence of endometriosis and, if present, performed rASRM empiric and algorithm assessment of disease stage. Reviewers had the option to select indeterminate at any given round if unable to diagnose or assess severity with reasonable accuracy. Prior to starting their online session, participating surgeons were asked to review the rASRM criteria for the staging of endometriosis. To avoid viewer fatigue, the system was programmed for up to 90 minutes of viewing at one sitting.

Statistical Analysis

The study population and expert reviewer characteristics were summarized using descriptive statistics. Three outcomes were derived to assess agreement between the operating surgeon's and expert reviewer's ratings: 1) a binary indicator of endometriosis as present or absent, 2) endometriosis staging by empiric categorization, and 3) endometriosis staging by rASRM weighted point score algorithm. Consistency of the MRI and confirmation of the histopathologic findings with the operating surgeon's endometriosis diagnosis (yes/no), consistent with operative real-time laparoscopic findings considered the gold standard,[9, 10] are also reported along with each diagnostic test's estimated sensitivity and specificity.

Inter-rater level agreement between the operating surgeon and the expert reviewer was calculated using kappa (κ) statistics for all review rounds: 1) digital images alone (n=148), 2) digital images + operative report (n=148), 3) digital images + operative report + MRI findings (n=36), 4) digital images + operative report + histopathological findings (n=67), and 5) digital images + operative report + MRI + histopathological findings (n=22). While the ENDO Study operating surgeon did not have access to the histopathologic findings nor routinely viewed the MRI, we wanted to assess how both of these added pieces of information might affect inter-rater agreement, as all are often made available for clinical, research, or medicolegal purposes. Furthermore, the level of reliance on endometriosis histopathology for an endometriosis diagnosis among expert raters was evaluated.

We calculated pair-wise Cohen's κ for endometriosis diagnosis and Cohen's weighted κ with squared weights for endometriosis staging. The κ statistic summarizes the rating data from 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 $\kappa = 0$, the agreement is no better than what would be obtained by chance alone. For our agreement analyses, we utilized Landis and Koch's[16] guidelines for interpreting κ statistics. Specifically, κ between 0.00 and 0.20 indicated slight agreement; κ between 0.21 and 0.40 denoted fair agreement; κ between 0.41 and 0.60 characterized moderate agreement; κ between 0.61 and 0.80 defined substantial agreement, and a value of κ greater than 0.80 equated to almost perfect agreement.

In the standard calculation for the Cohen's kappa, the p-value indicates whether the agreement is significantly different than zero, which can be achieved at relatively low values of kappa, particularly with a large enough sample size as was the case in our study, and thus not of great meaning.[16] In order to more meaningfully test the strength of agreement, we used Fleiss's guidelines where kappa > 0.75 is deemed excellent while <0.40 is deemed poor.[17] Thus for each κ, the p-value for testing H_0: $\kappa = 0.40$ versus H_1: $\kappa = 0.75$ also was estimated. Because both prevalence and bias play a part in determining the magnitude of the κ coefficient, we calculated prevalence-adjusted and bias-adjusted κ for assessing agreement for diagnosis due to unbalanced nature of the data, particularly when assessing agreement with added histopathologic review.[18] If a woman had an indeterminate diagnosis for endometriosis, then that woman was excluded from the final data analysis. Analyses were performed in SAS version 9.3 (SAS Institute, Cary, NC).

RESULTS

Three of the four reviewers had completed gynecologic laparoscopy fellowship training. Reviewers had an average of 15.0 years (IQR: 15.0–21.0) of clinical practice post-fellowship and 20.5 years (IQR: 17.0–23.5) post-residency; all were active practitioners, seeing on average 30.0 patients (IQR: 16.0–45.0) per week with gynecologic complaints or concerns. All reviewers had strong relevant publication records, having authored an average of 15.5 (IQR: 6.5–50.0) peer-reviewed research articles and 16.0 (IQR: 2.0–39.5) book chapters about endometriosis or surgical conditions of the pelvis. Study participants for this analysis (n=148) were on average 32.0 ± 6.7 years old, predominately non-Hispanic white (80.4%), married (83.7%), and nulliparous (51.4%). Primary reasons for laparoscopic surgery among the women in the study included pelvic pain (62.2%), pelvic mass (12.8%), menstrual irregularities (8.8%), infertility (7.4%), tubal ligation (6.1%), and fibroids (2.7%).

Mean percent agreement (MPA) between the experts' and operating surgeons' diagnosis was 83% (range: 80%–85%), but increased to 93% (range: 93%–95%) after viewing postoperative reports (**Table 1**). While neither MRI nor histopathologic findings alone improved agreement (mean: 90%, range: 86%–94%; mean: 85%, range: 81%–87%, respectively), agreement increased to 97% (range: 91%, 100%) after viewing both MRI and histopathologic findings in the subgroup for which these were both available (n=22). 2 × 2 tables indicated that expert reviewers assessed more participants as without endometriosis compared to the operating surgeon after additionally viewing operative and histopathologic (but not MRI) reports with between 7 to 10 women (13 to 20%) moving from a positive to negative diagnosis (Tables S1–S5).

Expert reviewers and operating surgeons also agreed, but to a lesser extent, on endometriosis staging by empiric assessment after viewing digital images (MPA: 49%, range: 44%–56%); the level of agreement increased after viewing operative reports (MPA: 60%, range: 58%–63%), but not after reviewing MRI and histopathologic findings (MPA: 47%, range: 46%–52%). MPA for endometriosis staging by algorithm based on the checklist was stronger for digital images (MPA: 57%, range: 52%–61%), and again increased after viewing operative reports (MPA: 72%, range: 68%–75%), but did not increase after viewing MRI and histopathologic findings (MPA: 57%, range: 27%–71%). Cross-classification tables showed that expert reviewers tended to diagnose more women with milder disease via empiric assessment compared to the operating surgeon after viewing operative images or operative images + operative reports (Tables S6–S7) but not after additionally viewing MRI or histopathology reports (Tables S8–S10). No clear differentiating patterns emerged for any rounds with rASRM checklist-assisted staging (Tables S11–S15).

Kappa statistics confirmed moderate to substantial agreement for endometriosis diagnosis (κ=0.67), and also for empiric (κ=0.60) and algorithm staging (κ=0.64). The level of agreement increased following review of operative reports (κ=0.88, 0.80, 0.85, respectively) (Table 1). There was no notable improvement in κ statistics with added MRI and a decrease in agreement with histopathologic findings. The vast majority of comparisons between the reviewers and operating surgeons for diagnosis and staging reached statistical significance at the preset hypothesis of κ=0.75 compared to the null hypothesis of κ=0.40, regardless of round (**Figures 1–3**).

Regarding diagnostic accuracy of MRI and histopathology, the MRI was consistent with the operating surgeon's report of endometriosis in 24 out of 36 women (67.7%), while the MRI did not detect an endometriosis diagnosis compared to the operating surgeon in the remaining 12 (33.3%) women, with the majority of these 12 women (n=9, 75%) with minimal or mild disease. There were no women for whom MRI detected endometriosis but the surgeon did not. Histopathologic reports confirmed the operating surgeon's assessment in 43 out of 67 women (63.2%), with the histopathologic report not detecting an endometriosis diagnosis compared to the operating surgeon in the remaining 24 (35.8%) women. The resulting sensitivity (proportion of women with the disease who are correctly identified by the test) and specificity (proportion of women without the disease who are correctly identified by the test) for MRI detected disease compared to the operating surgeon were 55.6% (95% confidence interval (CI); 35.3%, 74.5%) and 100% (95% CI: 66.2%, 100.0%), respectively. For histopathologically confirmed disease, sensitivity and specificity were 63.6% (95% CI, 59.9%, 75.1%) and 100% (95% CI; 16.7%, 100.0%), respectively.

DISCUSSION

Main Findings

In this study, an endometriosis diagnosis among experienced surgeons, using current diagnostic criteria, was shown to be consistent with expert review. Staging was more variable than diagnosis, particularly via empiric assessment versus from an algorithm based on the rASRM checklist. Reviewing operative reports in addition to digital images greatly improved agreement. This was not the case after additionally viewing MRI or more notably histopathologic findings, with expert reviewers designating up to 20% of women without endometriosis compared to the operating surgeon after viewing histopathology. Thus, in nearly 20% of cases, negative histopathology appears to overrule a visual diagnosis of endometriosis. Further study of the presentation, treatment, and outcomes of this group of visually diagnosed with endometriosis, but not confirmed by histopathology, will be informative. Improved staging systems of endometriosis with appropriate evaluation of reliability and validity based on current diagnostic criteria are warranted as is future research regarding the stage specific sensitivity of MRI; the potential sampling error in histologic findings; and the presentation, treatment, and outcomes of non-congruent visual and histologic diagnosis cases.

Strengths and Limitations

Our study had several major strengths, notably the inclusion of women with an array of clinical indications for laparoscopy being evaluated by a variety of operating surgeons and reassessed by expert reviewers from geographically diverse clinical centers. The findings may be particularly informative to researchers planning to assess the role and weight of positive and negative histopathology in the diagnosis of endometriosis. Still our study had important limitations largely reflecting its observational design and inability to have MRI and histology information on all women in keeping with the nature of clinical practice. Furthermore, it is important to note that laparoscopic surgeons did not have histology data available until after surgery. The extent to which this may impact the findings remains to be established with randomized trials where ethically possible. It is also important to keep in mind the relatively limited numbers of women undergoing MRI and histopathologic assessment that may result in imprecise kappa estimates. Larger studies among diverse surgical cohorts using additional and perhaps novel clinical assessment tools and/or biomarkers beyond laparoscopy to confirm diagnosis would be an important contribution to clinical practice and research based upon operative cohorts of women.

Interpretation

Our findings regarding diagnostic consistency complement and extend findings from previous research focusing on the variability of clinical diagnosis. Specifically, this literature includes one previous study focusing on the interobserver variability of endometriosis diagnosis[13] and two studies focusing on the variability of endometriosis staging.[13,14] Digital videotapes of laparoscopies from 3 patients (one with minimal, one with mild, and one with no endometriosis) were shown to 108 gynecologic surgeons for whom the surgeons assessed diagnosis and staging using the rASRM criteria.[13] Fifty-two percent of the reviewing surgeons agreed with the operating surgeon on the positive endometriosis diagnosis, 22% on the minimal staging, and 13% on the mild staging. While not directly comparable due to differences in study design (i.e., Buchweitz et al's study had 108 reviewers and 3 patients while our study had 4 reviewers and 148 patients), on average the four expert reviewers in our study agreed with the operating surgeon on a positive endometriosis diagnosis in 93% of the women, minimal staging in 83% of the women, and mild staging in 60% of the women (with rASRM checklist) after viewing operative images. While Buchweitz et al reported no differences by degree of training, our higher agreement may be due to our reviewers having extensive gynecological training versus their study comprising both board-certified specialists (43%) and residents (57%).[13] That the expert reviewers in our study diagnosed only 7% of women as having endometriosis when the surgeon diagnosed none, compared to the near 50% in the Buchweitz et al study, highlights the importance of ensuring adequate expertise when evaluating the need for endometriosis treatment among women, so as to minimize unnecessary surgical or medical therapies with potential side effects.[9,10,13]

The only other study to assess endometriosis staging agreement was done by Rock et al who found moderate agreement between surgeons (κ=0.44), with 96 out of 159 cases (60%) staged consistently by one blinded reviewer.[14] While staging agreement in this study more closely reflects our study findings after digital image review, comparison between the two studies is difficult given that the former study used the revised American Fertility Society (rAFS) classification system while the ENDO Study relied upon the rASRM criteria. Additionally, Rock et al's study included only women previously diagnosed with endometriosis and any women with images deemed inadequate to allow for classification were omitted from the analysis.[14]

Our study is the first to investigate whether MRI or histology findings influence endometriosis diagnosis and staging agreement relative to review of digitally recorded laparoscopies alone. Additionally viewing MRI findings did not greatly impact agreement. Given that the majority of women in the ENDO Study had minimal to mild disease and previous research showing MRI to be less accurate in detecting less severe disease,[19] our findings are not surprising. Unlike MRI, agreement substantially decreased after viewing histological findings with expert reviewers changing their assessment from a positive to negative diagnosis if findings did not confirm endometriosis diagnosis suggesting a preference toward histological confirmation. Most recent ESHRE guidelines suggest histopathologic confirmation as a useful diagnostic criterion.[10] Additionally, expert reviewers were more likely to diagnose more mild disease after viewing histology. Whether these patients truly did have less severe disease than originally assigned by operating surgeon or whether the biopsy failed to sample and/or detect disease is not clear and deserves future study regarding the preference demonstrated, as well as diagnostic criteria. Study of visually and histologically discordant cases presentations, management, and treatment outcomes will be informative.

We found comparable diagnostic accuracy between the MRI and laparoscopic findings in three studies, which reported sensitivities between 61% and 74%.[19–21] Like Stratton et al, we found no evidence that MRI overestimates disease, but that it is limited in its ability to identify and stage minimal or mild disease. The random sampling of those receiving MRI enhances the validity of MRI sensitivity. Sensitivity of histopathologic confirmation of surgically viewed endometriosis has shown wide variability in previous studies (30 to 86%).[19, 22–24] Our finding is in line with evidence indicating that while positive histopathology has been shown to confirm surgically viewed endometriosis (i.e., true positives), negative histopathology does not exclude it (i.e., false negatives).[9]

CONCLUSION

In summary, we found substantial agreement for diagnosis and staging of endometriosis between expert reviewers and the original operating surgeons after viewing digital images, and excellent agreement after additionally viewing the operative report. Our findings lend confidence that misclassification bias in diagnosis or staging on disease outcome is minimal. The higher agreement found with the rASRM algorithm compared to empiric assessment of staging warrants consideration of validated checklists for endometriosis research studies. Furthermore, our findings should alert consulting gynecologists who review operative images in order to make decisions on treatment options for endometriosis that operative reports/drawings and histopathology, but not necessarily MRI, will improve their ability to make sound judgments.

While experienced gynecologic surgeons appear to be consistent in their diagnoses, the consequences of incorrect or incomplete diagnoses on resultant therapy or disease course has not been sufficiently studied. Novel methods for facilitating endometriosis diagnostic procedures have been proposed, including the use of single or combined biomarkers.[25] Future classification systems that can provide consistency in the diagnosis and staging of endometriosis and better correlate with the clinical outcomes of pelvic pain and infertility will result in more accurate disease burden estimations as well as treatment efficacy assessments.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

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 and developing the online review system; and the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development ENDO Study database managers Christina Bryant and Jansen Davis for their help in extracting and tabulating the information from the online review system.

FUNDING

Supported 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, and 10001406-02).

REFERENCES

1. 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; 96:360–365. [PubMed: 21719000]

2. Rawson JM. Prevalence of endometriosis in asymptomatic women. J Reprod Med. 1991; 36:513–515. [PubMed: 1834839]

3. Giudice LC. Clinical practice. Endometriosis. N Engl J Med. 2010; 362:2389–2398. [PubMed: 20573927]

4. Balasch J, Creus M, Fábregues F, Carmona F, Ordi J, Martinez-Román S, et al. Visible and non-visible endometriosis at laparoscopy in fertile and infertile women and in patients with chronic pelvic pain: a prospective study. Hum Reprod. 1996; 11:387–391. [PubMed: 8671229]

5. Eskenazi B, Warner ML. Epidemiology of endometriosis. Obstet Gynecol Clin North Am. 1997; 24:235–258. [PubMed: 9163765]

6. Hsu AL, Sinaii N, Segars J, Nieman LK, Stratton P. Relating pelvic pain location to surgical findings of endometriosis. Obstet Gynecol. 2011; 118:223–230. [PubMed: 21775836]

7. Nnoaham KE, Webster P, Kumbang J, Kennedy SH, Zondervan KT. Is early age at menarche a risk factor for endometriosis? A systematic review and meta-analysis of case-control studies. Fertil Steril. 2012; 98:702–712. [PubMed: 22728052]

8. Practice Committee of the American Society for Reproductive M. Endometriosis and infertility: a committee opinion. Fertil Steril. 2012; 98:591–598. [PubMed: 22704630]

9. 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 Repro. 2005; 20:2698–2704.

10. Dunselman GA, Vermeulen N, Becker C, Calhaz-Jorge C, D'Hooghe T, De Bie B, et al. ESHRE guideline: management of women with endometriosis. Hum Reprod. 2014; 29:400–412. [PubMed: 24435778]

11. Wykes CB, Clark TJ, Khan KS. Accuracy of laparoscopy in the diagnosis of endometriosis: a systematic quantitative review. BJOG. 2004; 111:1204–1212. [PubMed: 15521864]

12. Weijenborg 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]

13. Buchweitz O, Wulfing P, Malik E. Interobserver variability in the diagnosis of minimal and mild endometriosis. Eur J Obstet Gynecol Reprod Biol. 2005; 122:213–217. [PubMed: 16219522]

14. 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]

15. Schliep KC, Stanford JB, Chen Z, Zhang Bo, Dorais JK, Johnstone EB, et al. Interrater and intrarater reliability in the diagnosis and staging of endometriosis. Obstet Gynecol. 2012; 120:104–112. [PubMed: 22914398]

16. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005; 37:360–3. [PubMed: 15883903]

17. Fleiss, JL.; Levin, B.; Palk, MC. Statistical methods for rates and proportions. 3rd edn.. John Wiley & Sons; New York: 2003.

18. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005; 85:257–268. [PubMed: 15733050]

19. Stratton P, Winkel C, Premkumar A, Chow C, Wilson J, Hearns-Stokes R, et al. Diagnostic accuracy of laparoscopy, magnetic resonance imaging, and histopathologic examination for the detection of endometriosis. Fertil Steril. 2003; 79:1078–1085. [PubMed: 12738499]

20. Tanaka YO, Itai Y, Anno I, Matsumoto K, Ebihara R, Nishida M. MR staging of pelvic endometriosis: role of fat-suppression T1-weighted images. Radiat Med. 1996; 14:111–116. [PubMed: 8827803]

21. Ha HK, Lim YT, Kim HS, Suh TS, Song HH, Kim SJ. Diagnosis of pelvic endometriosis: fat-suppressed T1-weighted vs conventional MR images. Am J Roentgenol. 1994; 163:127–131. [PubMed: 8010198]

22. Chatman DL, Zbella EA. Biopsy in laparoscopically diagnosed endometriosis. J Reprod Med. 1987; 32:855–857. [PubMed: 2963124]

23. Moen MH, Halvorsen TB. Histopathological confirmation of endometriosis in different peritoneal lesions. Acta Obstet Gynecol Scand. 1992; 71:337–342. [PubMed: 1326207]

24. Walter AJ, Hentz JG, Magtibay PM, Cornella JL, Magrina JF. Endometriosis: correlation between histopathological and visual findings at laparoscopy. Am J Obstet Gynecol. 2001; 184:1407–1411. discussion 1411–1413. [PubMed: 11408860]

25. Yang H, Zhu L, Wang S, Lang J, Xu T. Noninvasive diagnosis of moderate to severe endometriosis: the platelet-lymphocyte ratio cannot be a neoadjuvant biomarker for serum cancer antigen 125. J Minim Invasive Gynecol. 2015; 22:373–7. [PubMed: 23850516]

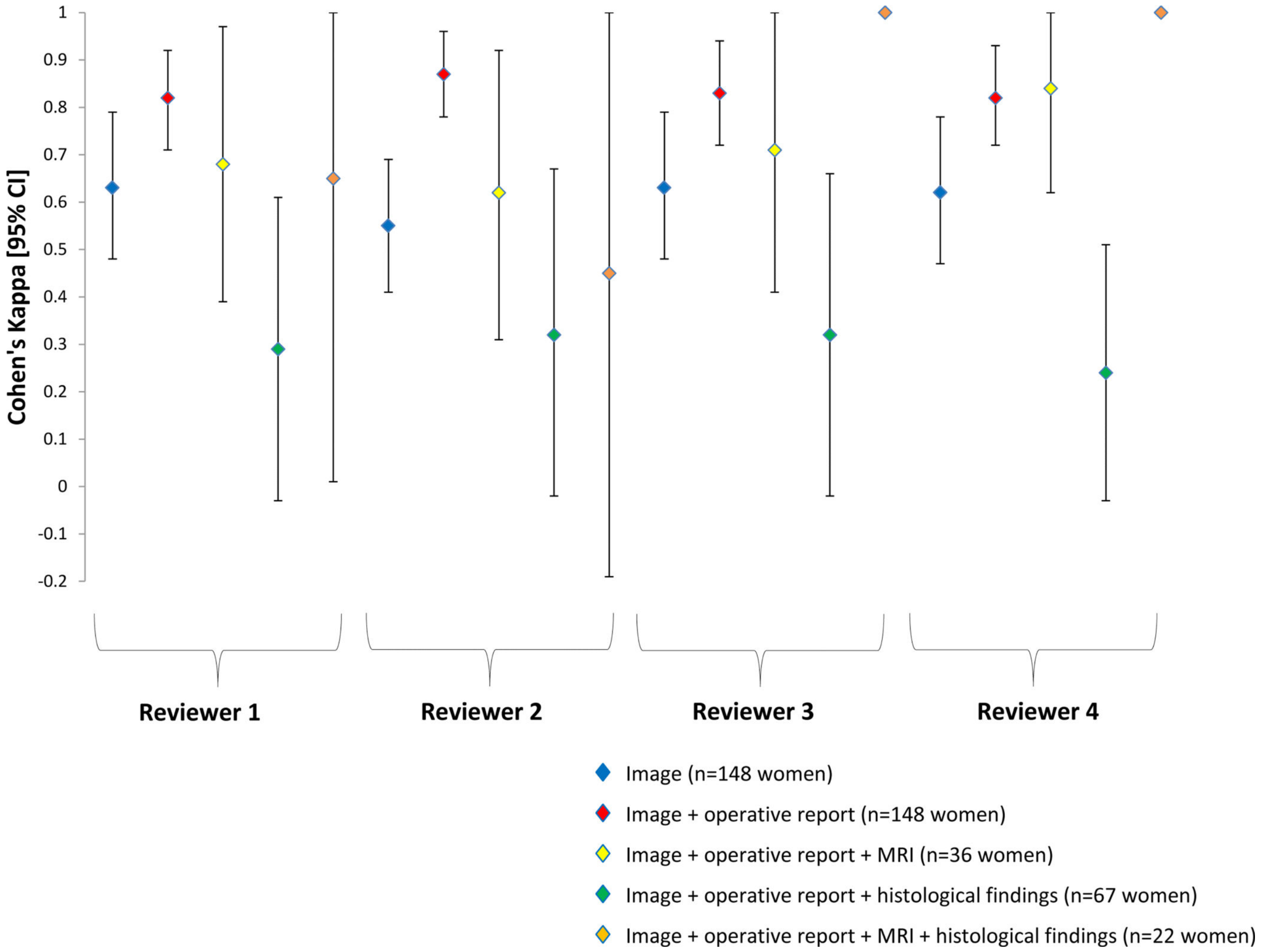

Figure 1. Endometriosis diagnosis

Unadjusted pair-wise Cohen's kappa (κ) and 95% CI of endometriosis diagnosis by operating surgeon and expert reviewer during real-time laparoscopy in the Endometriosis, Natural History, Diagnosis, and Outcomes Study, Salt Lake City, UT, USA, 2007–09. Prevalence and bias adjusted κ for Reviewers 1–4 after viewing images were 0.69, 0.60, 0.69, and 0.69 respectively; after images + operative report were 0.89, 0.89, 0.86, and 0.85 respectively; after images + operative report + MRI were 0.78, 0.72, 0.81, and 0.88; after images + operative report + histologic findings were 0.69, 0.75, 0.71, and 0.62; and after images + operative report + MRI histologic findings were 0.91, 0.82, 1.00, and 1.00. Note: No confidence intervals when perfect agreement (κ=1.0).

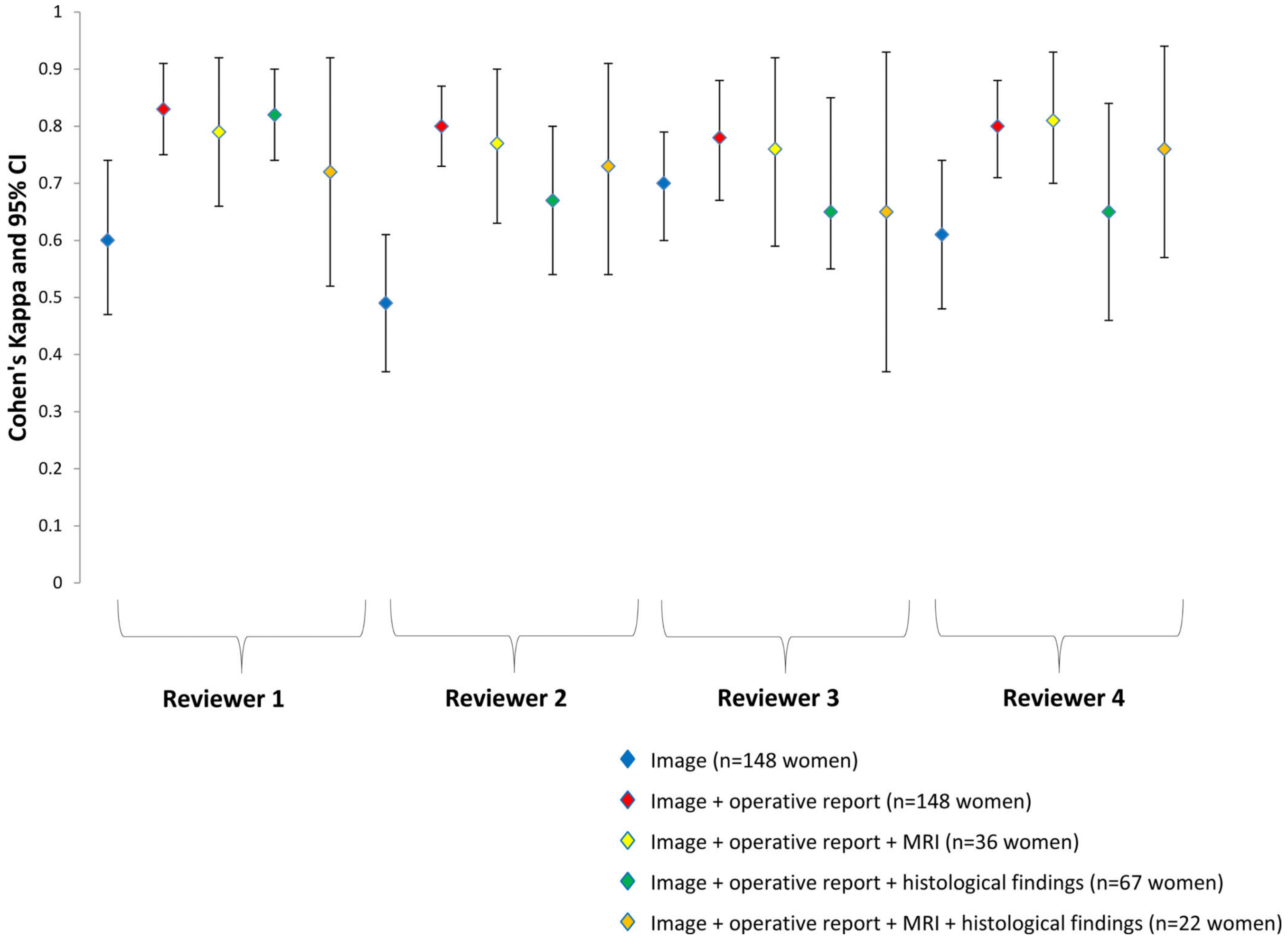

Figure 2. Endometriosis Staging, I-IV (Empiric Assessment)

Cohen's weighted kappa (κ) and 95% CI of endometriosis staging based on rASRM empiric assessment between operating surgeon and expert reviewer during real-time laparoscopy in the Endometriosis, Natural History, Diagnosis, and Outcomes Study, Salt Lake City, UT, USA, 2007–09.

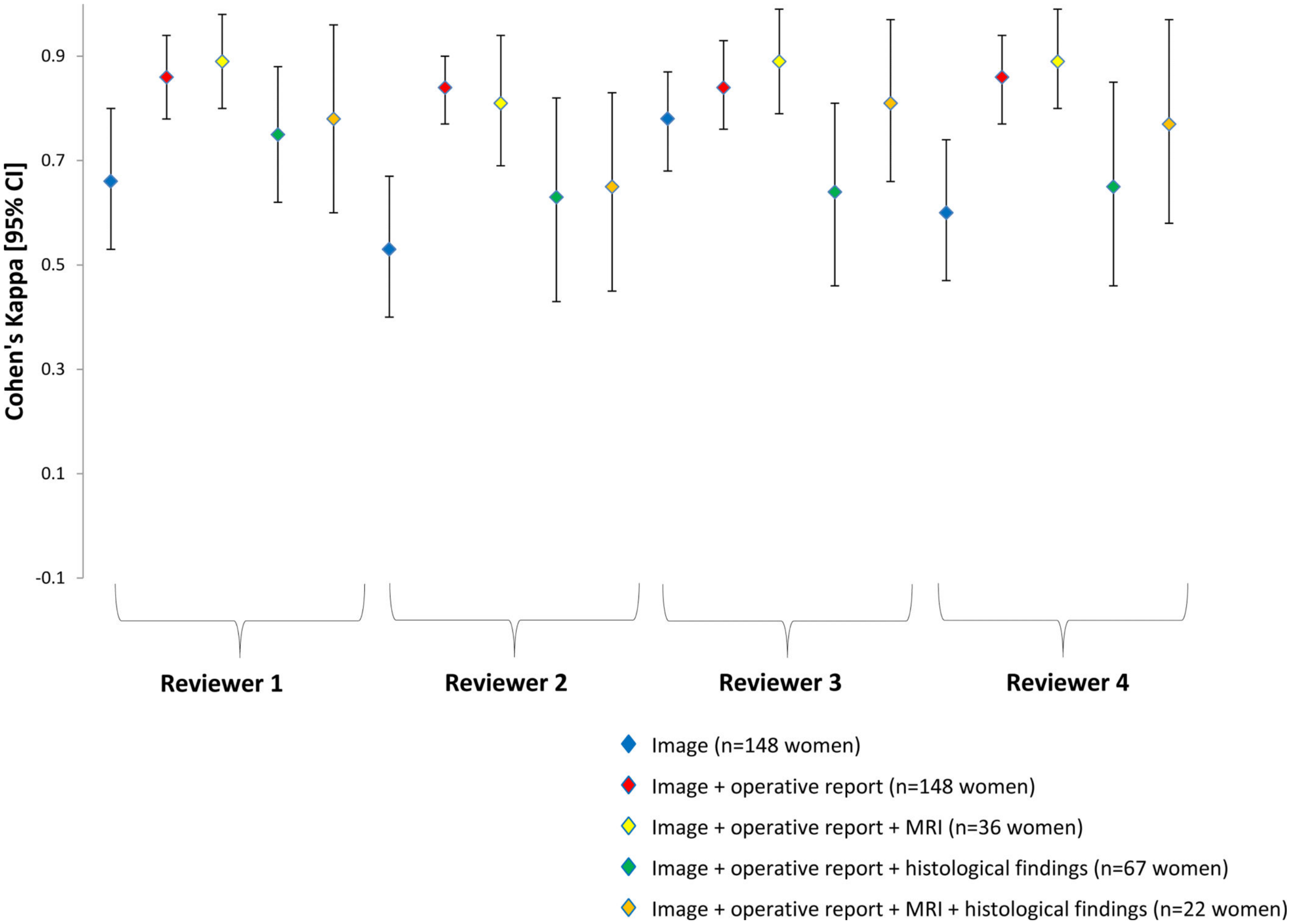

Figure 3. Endometriosis Staging, I-IV (Algorithm Assessment)

Cohen's weighted kappa (κ) and 95% CI of endometriosis staging based on rASRM algorithm assessment between operating surgeon and expert reviewer during real-time laparoscopy in the Endometriosis, Natural History, Diagnosis, and Outcomes Study, Salt Lake City, UT, USA, 2007–09.

Table 1

Comparability of endometriosis diagnosis and staging by operating surgeon and expert reviewer during real-time laparoscopy in the Endometriosis, Natural History, Diagnosis, and Outcomes Study, Salt Lake City, UT, USA, 2007–09.

Agreement	Digital Images (n=148)	Digital Images + Operative Report (n=148)	Digital Images + Operative Report + MRI Findings (n=36)	Digital Images + Operative Report + Histopathology Findings (n=67)	Digital Images + Operative Report + MRI + Histopathology Findings (n=22)
Endometriosis Diagnosis					
Mean % Agreement (Range)	83.4 (80.3, 84.6)	93.4 (92.7, 94.6)	89.9 (86.1, 93.9)	84.7 (80.8, 87.3)	96.7 (91.3, 100.0)
Mean Kappa[1] (Range)	0.67 (0.61, 0.69)	0.88 (0.85, 0.89)	0.80 (0.72, 0.88)	0.69 (0.62, 0.75)	0.93 (0.81, 1.00)
Endometriosis Staging, I-IV (Empiric Assessment)					
Mean % Agreement (Range)	48.5 (43.5, 55.6)	60.3 (58.1, 62.9)	51.1 (46.9, 57.6)	50.3 (47.2, 54.7)	47.2 (45.5, 52.4)
Mean Kappa (Range)	0.60 (0.49 0.70)	0.80 (0.78, 0.83)	0.78 (0.76, 0.81)	0.70 (0.65, 0.82)	0.72 (0.65, 0.76)
Endometriosis Staging, I-IV (Algorithm Assessment)					
Mean % Agreement (Range)	56.7 (52.1, 60.6)	71.8 (68.3, 74.6)	69.2 (54.5, 76.7)	51.1 (44.9, 56.6)	56.5 (27.2, 71.4)
Mean Kappa (Range)	0.64 (0.53, 0.78)	0.85 (0.84, 0.86)	0.87 (0.81, 0.89)	0.67 (0.63, 0.75)	0.75 (0.65, 0.81)

[1] Prevalence and bias adjusted