Diagnostic accuracy of a novel endometriosis staging system: an external validation study

Jason Mak[1], Allie Eathorne[2], Mathew Leonardi[3], Mercedes Espada[4], Shannon Reid[5], Jose Zanardi[6], Cansu Uzuner[7], Rodrigo Rocha[8], Mike Armour[9], and George Condous[8]

[1]Affiliation not available
[2]Medical Research Institute of New Zealand
[3]McMaster University
[4]Blue Mountains District ANZAC Memorial Hospital
[5]University of Western Sydney
[6]Fecundita Endometriosis and Infertility Clinic
[7]Nepean Blue Mountains Local Health District
[8]Nepean Hospital
[9]Western Sydney University

May 2, 2022

Abstract

Objective To externally validate the "2021 AAGL Endometriosis Classification" staging system. Design Retrospective, diagnostic accuracy study Setting Multicentre Population or Sample Two hundred and seventy-two endometriosis patients (January 2016 - October 2021) Methods Three independent observers analysed coded surgical data to assign an AAGL surgical stage (1 to 4) as the index test, and surgical complexity level (A to D) as the reference standard. Main Outcome Measures The diagnostic accuracy of each AAGL stage to predict corresponding AAGL surgical complexity level was determined. Receiver operating characteristic curves used to determine the accuracy of cut off points used in the AAGL staging system to discriminate between surgical complexity levels. Results 272 cases were analysed. Diagnostic accuracy (sensitivity, specificity, PPV and NPV) for three observers were: stage 1 to predict level A 97.9-98.7%, 60.2-64.2%, 75.0-76.9%, and 96.3-97.5%; stage 2 to predict level B 26.1-30.4%, 93.2-95.6%, 26.3-35.3%, and 92.9-93.6%; stage 3 to predict level C 7.5-10.0%, 93.8-94.8%, 33.3-42.1%, and 70.9-71.5%; stage 4 to predict level D 90.-95.0%, 90.1-91.7% &, 41.9-47.5%, and 99.1-99.6%. For three observers AUROC for A vs B/C/D (cut-point 9) 0.75-0.88, A/B vs C/D (cut-point 16) 0.81 and A/B/C vs D (cut-point 22) 0.95-0.96. Conclusions This external validation study demonstrates that the AAGL Endometriosis Classification performs poorly overall for the prediction of surgical complexity. The results from this external validation study suggest that the system in its current form is not generalizable to all endometriosis patients and should be reviewed before its universal implementation. Funding Nil Keywords Endometriosis, staging, laparoscopy

Title:

Diagnostic accuracy of a novel endometriosis staging system: an external validation study

Authors:

Jason Mak[1,2], Allie Eathorne[3], Matthew Leonardi[4], Mercedes Espada[5,6], Shannon Reid[7,8], Jose Vitor Zanardi[9,10], Cansu Uzuner[1,2], Rodrigo Rocha[1], Mike Armour[11], George Condous[1,,6]

Affiliations:

1. Acute Gynaecology, Early Pregnancy, and Advanced Endosurgery Unit, Nepean Hospital, Sydney, Australia
2. Sydney Medical School Nepean, The University of Sydney, Sydney, Australia
3. Medical Research Institute of New Zealand, Wellington, New Zealand
4. Department of Obstetrics and Gynecology, McMaster University, Hamilton, Canada
5. Department of Obstetrics and Gynaecology, Blue Mountains District ANZAC Memorial Hospital, Sydney, Australia
6. OMNI Ultrasound & Gynaecological Care, Sydney, Australia
7. Department of Obstetrics and Gynaecology, Western Sydney University, Sydney, Australia
8. Department of Obstetrics and Gynaecology, Campbelltown Private Hospital, Sydney, Australia
9. Fecundita Endometriosis and Infertility Clinic, Ribeirao Preto, Brazil
10. Ribeirao Preto School of Medicine, University of Sao Paulo, Brazil
11. NICM Health Research Institute, Western Sydney University, Sydney, Australia

Corresponding author

Dr Jason Mak

Acute Gynaecology, Early Pregnancy, and Advanced Endoscopy Surgery Unit, Sydney Medical School Nepean, Nepean Hospital, University of Sydney, Sydney, 2747, NSW, Australia

Telephone: +61 0451877943; Facsimile: +61 2 4734 4887

e-mail: jason.mak@sydney.edu.au

Short running title:

AAGL endometriosis staging system external validation

Abstract

Objective

To externally validate the "2021 AAGL Endometriosis Classification" staging system.

Design

Retrospective, diagnostic accuracy study

Setting

Multicentre

Population or Sample

Two hundred and seventy-two endometriosis patients (January 2016 - October 2021)

Methods

Three independent observers analysed coded surgical data to assign an AAGL surgical stage (1 to 4) as the index test, and surgical complexity level (A to D) as the reference standard.

Main Outcome Measures

The diagnostic accuracy of each AAGL stage to predict corresponding AAGL surgical complexity level was determined. Receiver operating characteristic curves used to determine the accuracy of cut off points used in the AAGL staging system to discriminate between surgical complexity levels.

Results

272 cases were analysed. Diagnostic accuracy (sensitivity, specificity, PPV and NPV) for three observers were: stage 1 to predict level A 97.9-98.7%, 60.2-64.2%, 75.0-76.9%, and 96.3-97.5%; stage 2 to predict level

B 26.1-30.4%, 93.2-95.6%, 26.3-35.3%, and 92.9-93.6%; stage 3 to predict level C 7.5-10.0%, 93.8-94.8%, 33.3-42.1%, and 70.9-71.5%; stage 4 to predict level D 90.-95.0%, 90.1-91.7% &, 41.9-47.5%, and 99.1-99.6%. For three observers AUROC for A vs B/C/D (cut-point 9) 0.75-0.88, A/B vs C/D (cut-point 16) 0.81 and A/B/C vs D (cut-point 22) 0.95-0.96.

Conclusions

This external validation study demonstrates that the AAGL Endometriosis Classification performs poorly overall for the prediction of surgical complexity. The results from this external validation study suggest that the system in its current form is not generalizable to all endometriosis patients and should be reviewed before its universal implementation.

Funding

Nil

Keywords

Endometriosis, staging, laparoscopy

Tweetable abstract

AAGL endometriosis staging system ext validation: 272 cases. Poor diagnostic accuracy (stage to predict surgical complexity)

Main Text

Introduction

Endometriosis staging tools have undergone forty years of evolution and multiple iterations, however none has yet gained universal acceptance. The proposed utility of an endometriosis staging tool is also not universally agreed. It is acknowledged that endometriosis staging has an important role in stratifying the disease for research purposes (1), predicting surgical complexity and potentially having a utility with reimbursement (2). Ideally, assignment of an endometriosis stage might be useful in communicating clinically relevant disease severity. To date this has not been achieved. Survey data suggests that most users of existing endometriosis staging tools site a lack of clinical relevance as the main limitation, and would welcome a new tool (3).

The first attempt at classifying endometriosis was published in Lockyer's book "Fibroids and allied tumors", in 1918 (4). Since then, the most well-known and widely utilised staging system has been the revised American Society for Reproductive Medicine (rASRM) classification. The system was first published by the ASRM in 1979 (5) and has undergone two revisions, the latest in 1996 (6, 7). Endometriosis is notorious for poor correlation between disease burden and symptomatology. This phenomenon has made it difficult to develop a classification system that predicts clinical outcomes relevant to the patient, which is one of many criticisms of the rASRM staging system (8). It does not correlate with pain, quality of life, fertility or treatment outcomes (8, 9). In addition, it has been criticised for failing to address deep endometriosis and retroperitoneal structures (8, 9). It is arguably time-consuming and cumbersome to use. It's usefulness is further challenged by the fact that poor interobserver variability has been demonstrated (10).

The three best known attempts at improving the rASRM system have been #Enzian, the Endometriosis Fertility Index (EFI) and the 2021 AAGL Endometriosis Classification. The #Enzian classification system, most recently updated in 2021 (11) after several iterations (12, 13), was originally designed to complement the rASRM system and address deep endometriosis (14). The latest edition is more comprehensive and designed to stand alone encompassing both deep disease, superficial endometriosis and adhesions (11). #Enzian does not result in a global severity stage. Rather, it maps disease in seven separate anatomical domains. It is therefore difficult to quantifiably compare #Enzian to any staging tool.

The EFI is a scoring tool that aims to predict pregnancy rates in individuals with endometriosis (15). It incorporates three components: surgical findings in the form of the rASRM, a functional score of the tubes and ovaries and clinical factors such as age, duration of infertility and previous pregnancy. A recent metanalysis of seventeen studies found the EFI performs well at predicting spontaneous pregnancy rates (16). The tool has also demonstrated good inter-observer agreement (17). Most disagreements in EFI occurred on account of differences in the rASRM score component, suggesting the tool might be amenable to improvement by replacing rASRM with another global staging system.

The "2021 AAGL Endometriosis Classification" staging system, henceforth referred to as the AAGL system, like its predecessor the rASRM, is a points-based staging system (2). A table of anatomical and pathological laparoscopic findings are listed which generate corresponding points, directly proportionate to disease severity. The total point score is then applied to thresholds that determine surgical complexity stages 1 to 4. A large prospective trial demonstrated a high concordance between the AAGL stage and surgical complexity, superior to the rASRM when compared head-to-head (2). Correlation with pain and fertility was also demonstrated, again, superior to the rASRM (2). To our knowledge, this staging system has yet to be externally validated in terms of its stated purpose as a diagnostic tool for predicting surgical complexity. Our objective is to externally validate the diagnostic test performance of the AAGL system.

Methods

Study design

We performed a multicentre retrospective diagnostic accuracy study on women with suspected endometriosis. The study was approved by the Nepean Blue Mountains Local Health District ethics committee; 2022/ETH00000.

Participants

Information on the 272 women included in the current study were from two databases, published in three separate studies, which described the use of sonovaginography for the prediction of surgical complexity (18-20). Participants in the original studies were seen between January 2016 and October 2021 in Sydney, Australia. They underwent surgery with one of six minimally invasive gynaecological surgeons (MIGS), all with the highest level of surgical ability, as per the Royal Australian New Zealand College of Obstetricians and Gynaecologists/Australasian Gynaecological Endoscopy & Surgery Society (RANZCOG/AGES) (21). Each participant had a systematic visual inspection of the pelvis, upper abdomen, and appendix. The databases used in this paper therefore contained comprehensive, coded data mapping the location and morphology of disease, as well as the surgical procedures performed for each case. Surgical findings were arranged under the columns of anterior peritoneal, anterior deep, posterior peritoneal, posterior deep, lateral and adnexal pathology. Under these columns pathology was coded and mapped under 66 subheadings. In addition, the presence of hydroureter and the presence and extent of pouch of Douglas obliteration (partial or complete) was described. Likewise, surgical procedures were listed including ureterolysis and the type of bowel surgery performed. The databases from the three studies were interrogated to ensure there were no cases with incomplete data. The participants were then de-identified.

Inclusion criteria from the original studies were women of reproductive age, and either a history of chronic pelvic pain, or a history of endometriosis, or both. Exclusion criteria were women with suspected malignancy, pregnancy, premature ovarian failure, menopause, and nonsurgical management. Additional exclusion criteria for this study were patients with incomplete data. All women included in the databases had previously consented for their de-identified surgical data to be used in research.

Test Methods

The index test was AAGL stage, and the reference standard was AAGL surgical complexity level. Based on the coded surgical data, each participant could be reviewed and AAGL score, stage and also AAGL surgical complexity level could be retrospectively apportioned in the same manner as a patient being staged in real time at laparoscopy. Where detail on the size of lesions was missing, the maximum number of possible points for that region was used to calculate the AAGL score. Maximum scores were used so as not to underestimate the severity of disease.

The de-identified database was presented to three expert observers who were either MIGS or fellows in their final year of MIGS training. Firstly, the three observers were asked to allocate an AAGL surgical complexity level (A to D) for each case, as defined in the paper by*Abrão et al (2)* . A single, reference AAGL surgical complexity level allocation was then developed by consensus. Next, the three observers were blinded to the AAGL surgical complexity level and asked to allocate an AAGL stage for each case. Staging was allocated twice. For the first allocation run, observers were left to interpret the AAGL staging tool and schematic in the paper by *Abrão et al*independently (2). The three observers then met, discussed the tool and developed consensus rules of interpretation. The three observers then performed a second staging allocation run for each case, blinded to the first. The stages from the second allocation run were used in the final analysis, to optimise interobserver agreement.

Analysis

We calculated the diagnostic performance of each observer to predict level of surgical complexity for each stage, i.e. AAGL stage 1 to predict level A, AAGL stage 2 to predict level B, AAGL stage 3 to predict level C and AAGL stage 4 to predict level D. Data were analysed to determine the kappa and weighted kappa scores, accuracy, sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood ratio, and negative likelihood ratio, with 95% confidence intervals. The AAGL system uses a cumulative point score schema, and stage is determined by score thresholds. The paper by*Abrão et al.* describes logistic regression to determine the point score thresholds defining stages 1-4, that would most accurately predict skill levels A–D (2). Stage 1 was determined to be 0 – 8 points, stage 2 was 9 to 15 points, stage 3 was 16 to 21 points and stage 4 was above 21 points. We tested our dataset in the same manner: area under the receiver operating characteristic curves (AUROC) were used to determined overall performance of A vs B/C/D (for a threshold of 8), A/B vs C/D (for a threshold of 15) & A/B/C vs D (for a threshold of 21), for each observer.

Continuous data were summarised by mean and standard deviation, median and interquartile range (25$^{\text{th}}$ to 75$^{\text{th}}$ percentile), and minimum to maximum. Categorical data were summarised by counts and proportions expressed as percentages. Ordinal data are described by cross-tabulation and summarised as described for continuous data.

Funding

There was no funding applied for, or received for this research.

Results

272 patients were included in the final analysis. The database used in the *Leonardi et al.* (19) and *Espada et al.* (18) papers containing 204 patients was reduced to 194 after incomplete data was identified in 10 cases. The database used in the *Rao et al.* (20) paper included 78 patients. All were complete and therefore included.

Summary data is presented in Tables 1 and 2. Overall, AAGL stage by three observers accurately predicted the corresponding AAGL surgical complexity level in 175 – 180 of the 272 cases (64.3 – 66.2%). The overall performance of the AAGL system in terms of kappa and weighted kappa scores, accuracy, sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood ratio, and negative likelihood ratio to predict AAGL level of laparoscopic surgical complexity is summarised in Tables 3, 4 and 5. Best performance of three observers for sensitivity, specificity, PPV and NPV (95% CI's) for stage 1 to predict complexity level A was 98.7% (96.8 to 100.5), 64.2% (55.8 to 72.8), 77.0% (71.0 to 82.9) and 97.5% (94.1 to 100.9) respectively. For stage 2 to predict level B was 30.4% (11.6 to 49.2), 95.6% (93.0 to 98.1), 35.3% (12.6 to 58.0) and 93.5% (90.5 to 96.6) respectively. For stage 3 to predict level C was 10.0% (3.4 to 16.5), 94.8% (91.6 to 97.9), 42.% (19.9 to 64.3) and 71.5% (65.9 to 77.1) respectively. For stage 4 to predict level D was

95.0% (85.4 to 104.5), 91.7% (88.2 to 95.1), 47.5% (32.0 to 63.0) and 99.6% (98.7 to 100.4) respectively. The performance of score thresholds 8, 15 and 21 for predicting corresponding skill levels (A – D) is reported in Table 6, and corresponding ROC curves are shown in Figure 1.

Discussion

Main findings

Stage 1 performed reasonably well at predicting surgical complexity level A, with high sensitivity and NPV, but moderate specificity and PPV. The intermediate stages 2 and 3 performed poorly for predicting corresponding surgical complexity levels. Stage 4 had poor PPV for predicting surgical complexity level D. Pre-determined staging thresholds performed well at discerning skill level A/B/C versus D (stage 4) but low specificity for A versus B/C/D and A/B/C versus D (stages 1, 2 and 3).

Strengths and Limitations

This retrospective study used descriptive data rather than laparoscopic images to simulate endometriosis staging, and therefore might not be as representative as real time intraoperative staging. A prospective study where scoring and staging is performed in the usual contemporaneous fashion would overcome this. The three assessors in this study were either MIGS or fellows in their final year of MIGS training, and data were obtained from tertiary pelvic pain referral centres. Selection bias has been identified as a challenge in endometriosis research (22), whereby sample populations recruited to studies are not always representative of the 11% (23) of the female population that suffer from endometriosis. These findings therefore might not be generalizable to general gynaecology clinics and generalist obstetricians and gynaecologists who might use this tool, for example.

A strength of our study was the use of a consensus process to reduce heterogeneity between assessors interpreting the AAGL system. The rationale for this was that the authors identified several potential areas of ambiguity within the tool. While this improved consistency between the three assessors in this experiment, it does not necessarily represent the real-world application of endometriosis staging tools. In addition, the fact that three observers were used and the best of three results was considered for discussion demonstrates that all due effort was made assess the AAGL system fairly. Bias was mitigated by blinding between observers, and also between the staging and skill level allocations.

Interpretation

The AAGL system accurately predicted surgical complexity level in 66.2% of cases, which is comparable to the 69.2% found in the original paper (2). In our study the overall agreement between AAGL stage and AAGL complexity level was weak, as quantified by a weighted kappa score of 0.38 – 0.42 across the three observers. This was low, compared with 0.621 in the original study (2), which suggested moderate agreement. Stage 1 performed reasonably well at predicting skill level A and this was consistent across the three observers, however the remaining stages 2, 3 and 4 did not correlate well.

The pre-specified AAGL cut-points had reasonably high specificity for discerning skill level A/B/C versus D (stage 4) but low specificity for A versus B/C/D and A/B/C versus D (lower levels). When AUCROC data in this external validation are directly compared to those reported in the paper by in the original paper (2), the results are less robust. For A versus B/C/D, AUCROC in the original paper was 0.98, and in our analysis, it was lower at 0.75 to 0.89. For A/B versus C/D, AUCROC in the original paper was 0.95, and in our analysis, it was lower at 0.81. For A/B/C versus D, AUCROC in the original paper was 0.91, and in our analysis, it was higher at 0.95 to 0.96. This may reflect the fact that in the original paper, regression analysis was used to identify optimal cut points for that particular dataset, so the performance would therefore be expected to be less promising when externally validated. Poor diagnostic accuracy for levels 2, 3 and 4 and lower than previously reported AUCROC results in our dataset suggest that the AAGL staging tool is not be generalizable in its current form.

While stage 4 had a low PPV for predicting surgical complexity level D (47.5%), the specificity (91.7%) and

PPV (99.57 %) were high. This demonstrates that stage 4 performs well at ruling out those without lower surgical complexity levels. The AUROC for stage 4 to discriminate level D from levels A/B/C was high at 0.95, which confirmed this finding. These results suggest that the tool might be useful for surgical planning, although if the stage can only be determined intraoperatively, the utility of this is limited.

Conclusion

The disease entity endometriosis is enigmatic and presents a major challenge to developers of potential staging systems. This external validation study demonstrates that the AAGL 2021 endometriosis classification is not generalisable. One of the key strengths of the tool is for high stage disease to distinguish a high level of surgical complexity. While this is of limited utility when endometriosis is diagnosed intraoperatively, these findings suggest that a corresponding pre-operative endometriosis staging system might provide the ability to triage and appropriately plan for the anticipated level of surgical complexity. This finding is consistent with the performance of existing ultrasound-based endometriosis staging tools, particularly with higher stage disease (24). Despite this, survey data suggests that both patients (25) and gynaecologists (26) have limited awareness of non-surgical endometriosis diagnosis.

In the ongoing evolution of endometriosis staging and the search for a universally accepted tool, the AAGL system has some key weaknesses and important strengths. We support the assertion that the ideal staging system should incorporate both a pre-operative and intra-operative arm, predict surgical complexity, correlate well with clinical outcomes relevant to the patient (27) and demonstrate internal and external validity. A prospective external validation of the recently published AAGL system should follow, including a head-to-head comparison with existing and emerging endometriosis staging systems. Standardized surgical data collection would be advisable, as per the Consensus on Recording Deep Endometriosis Surgery statement (28).

Acknowledgements

Nil

Disclosure of Interests

ML reports grants from OZWAC, Endometriosis Australia, AbbVie, CanSAGE, MRFF, HHS; honoraria for lectures/writing from GE Healthcare, Bayer, AbbVie, TerSera, consulting fees from Imagendo, outside the submitted work. JVZ reports honoraria for lectures from Ethicon/Johnson & Johnson, Merck Sharp & Dohme and Libbs Farmacêutica. MA is chair of the Endometriosis Australia research committee, chair of the Endometriosis Australia Clinical Advisory Committee, and has received funding from Metagenics outside the submitted work. GC reports grants from Endometriosis Australia and MRFF; honoraria for lectures from GE Healthcare and Imperial College London. All other authors report nothing to disclose.

Contribution to authorship

All authors contributed to the conception and planning of the study and editing of the manuscript. ML, ME, SR, JM and GC contributed cases to the final database. JM, CU and ME were the observers. AE performed the statistical analysis. JM coordinated the project and was the primary author of the manuscript.

Ethics approval

This research was ethically approved by the Nepean Blue Mountains Local Health District Human Research Ethics Committee under reference number 2022/ETH00000

Funding

Nil

References

1. Leonardi M, Horne AW. Important concepts to consider in endometriosis surgical trials. Bjog. 2020;127(12):1536.

2. Abrao MS, Andres MP, Miller CE, Gingold JA, Rius M, Neto JS, et al. AAGL 2021 Endometriosis Classification: An Anatomy-based Surgical Complexity Score. J Minim Invasive Gynecol. 2021;28(11):1941-50.e1.

3. Zondervan KT, Missmer S, Abrao MS, Einarsson JI, Horne AW, Johnson NP, et al. Endometriosis Classification Systems: An International Survey to Map Current Knowledge and Uptake. J Minim Invasive Gynecol. 2022.

4. Lockyer C. Fibroids and allied tumours (Myoma and Adenomyoma). London: MacMillan and Co; 1918.

5. Classification of endometriosis. The American Fertility Society. Fertil Steril. 1979;32(6):633-4.

6. Revised American Society for Reproductive Medicine classification of endometriosis: 1996. Fertil Steril. 1997;67(5):817-21.

7. Revised American Fertility Society classification of endometriosis: 1985. Fertil Steril. 1985;43(3):351-2.

8. Haas D, Shebl O, Shamiyeh A, Oppelt P. The rASRM score and the Enzian classification for endometriosis: their strengths and weaknesses. Acta Obstet Gynecol Scand. 2013;92(1):3-7.

9. Andres MP, Borrelli GM, Abrão MS. Endometriosis classification according to pain symptoms: can the ASRM classification be improved? Best Pract Res Clin Obstet Gynaecol. 2018;51:111-8.

10. Buchweitz O, Wülfing P, Malik E. Interobserver variability in the diagnosis of minimal and mild endometriosis. Eur J Obstet Gynecol Reprod Biol. 2005;122(2):213-7.

11. Keckstein J, Saridogan E, Ulrich UA, Sillem M, Oppelt P, Schweppe KW, et al. The #Enzian classification: A comprehensive non-invasive and surgical description system for endometriosis. Acta Obstet Gynecol Scand. 2021;100(7):1165-75.

12. Montanari E, Dauser B, Keckstein J, Kirchner E, Nemeth Z, Hudelist G. Association between disease extent and pain symptoms in patients with deep infiltrating endometriosis. Reprod Biomed Online. 2019;39(5):845-51.

13. Tuttlies F, Keckstein J, Ulrich U, Possover M, Schweppe KW, Wustlich M, et al. [ENZIAN-score, a classification of deep infiltrating endometriosis]. Zentralbl Gynakol. 2005;127(5):275-81.

14. Haas D, Wurm P, Shamiyeh A, Shebl O, Chvatal R, Oppelt P. Efficacy of the revised Enzian classification: a retrospective analysis. Does the revised Enzian classification solve the problem of duplicate classification in rASRM and Enzian? Arch Gynecol Obstet. 2013;287(5):941-5.

15. Adamson GD, Pasta DJ. Endometriosis fertility index: the new, validated endometriosis staging system. Fertil Steril. 2010;94(5):1609-15.

16. Vesali S, Razavi M, Rezaeinejad M, Maleki-Hajiagha A, Maroufizadeh S, Sepidarkish M. Endometriosis fertility index for predicting non-assisted reproductive technology pregnancy after endometriosis surgery: a systematic review and meta-analysis. Bjog. 2020;127(7):800-9.

17. Tomassetti C, Bafort C, Meuleman C, Welkenhuysen M, Fieuws S, D'Hooghe T. Reproducibility of the Endometriosis Fertility Index: a prospective inter-/intra-rater agreement study. Bjog. 2020;127(1):107-14.

18. Espada M, Leonardi M, Aas-Eng K, Lu C, Reyftmann L, Tetstall E, et al. A Multicenter International Temporal and External Validation Study of the Ultrasound-based Endometriosis Staging System. J Minim Invasive Gynecol. 2021;28(1):57-62.

19. Leonardi M, Espada M, Choi S, Chou D, Chang T, Smith C, et al. Transvaginal Ultrasound Can Accurately Predict the American Society of Reproductive Medicine Stage of Endometriosis Assigned at Laparoscopy. J Minim Invasive Gynecol. 2020;27(7):1581-7.e1.

20. Rao T, Condous G, Reid S. Ovarian Immobility at Transvaginal Ultrasound: An Important Sonographic Marker for Prediction of Need for Pelvic Sidewall Surgery in Women With Suspected Endometriosis. J Ultrasound Med. 2022;41(5):1109-13.

21. RANZCOG. Guidelines for performing gynaecological endoscopic procedures. Melbourne: Royal Australian and New Zealand College of Obstetricians and Gynaecologists; 2019.

22. De Graaff AA, Dirksen CD, Simoens S, De Bie B, Hummelshoj L, D'Hooghe TM, et al. Quality of life outcomes in women with endometriosis are highly influenced by recruitment strategies. Hum Reprod. 2015;30(6):1331-41.

23. Rowlands IJ, Abbott JA, Montgomery GW, Hockey R, Rogers P, Mishra GD. Prevalence and incidence of endometriosis in Australian women: a data linkage cohort study. Bjog. 2021;128(4):657-65.

24. Menakaya U, Reid S, Lu C, Gerges B, Infante F, Condous G. Performance of ultrasound-based endometriosis staging system (UBESS) for predicting level of complexity of laparoscopic surgery for endometriosis. Ultrasound Obstet Gynecol. 2016;48(6):786-95.

25. Menakaya UA, Hannedege B, Infante F, Lanzarone V, Adno A, Johnson NP. Awareness and utilisation of advanced gynaecological ultrasound in the preoperative work-up of women planning surgery for endometriosis: A survey of RANZCOG fellows and trainees. Aust N Z J Obstet Gynaecol. 2021;61(4):563-8.

26. Leonardi M, Rocha R, Tun-Ismail AN, Robledo KP, Armour M, Condous G. Assessing the knowledge of endometriosis diagnostic tools in a large, international lay population: an online survey. Bjog. 2021;128(13):2084-90.

27. Espada M, Leonardi M, Reid S, Condous G. Regarding "AAGL 2021 Endometriosis Classification: An Anatomy-based Surgical Complexity Score". J Minim Invasive Gynecol. 2022;29(3):449-50.

28. Johnson NP, Hummelshoj L, Adamson GD, Keckstein J, Taylor HS, Abrao MS, et al. World Endometriosis Society consensus on the classification of endometriosis. Hum Reprod. 2017;32(2):315-24.

Hosted file

`BJOG AAGL ext val tables 1.5.22.docx` available at https://authorea.com/users/480083/articles/567642-diagnostic-accuracy-of-a-novel-endometriosis-staging-system-an-external-validation-study

Hosted file

`BJOG AAGL ext val Figures 1.5.22.docx` available at https://authorea.com/users/480083/articles/567642-diagnostic-accuracy-of-a-novel-endometriosis-staging-system-an-external-validation-study