Article type : Main research article

TITLE PAGE:

Reproducibility of the Endometriosis Fertility Index: a prospective inter/intra-rater agreement study

Authors: C. Tomassetti[1, 2]*, C. Bafort[1], C. Meuleman[1, 2], M. Welkenhuysen[1], S. Fieuws[3], T. D'Hooghe[2]

[1] Department of Obstetrics and Gynaecology, Leuven University Fertility Centre, University Hospitals Leuven, Herestraat 49, Leuven 3000, Belgium

[2] Department of Development and Regeneration, KU Leuven, Herestraat 49, Leuven 3000, Belgium

[3] Department of Public Health, Interuniversity Centre for Biostatistics and Statistical Bioinformatics, KU Leuven, Kapucijnenvoer 35, Leuven 3000, Belgium.

* Correspondence address. E-mail: carla.tomassetti@uzleuven.be

Shortened running title: Reproducibility of the EFI

ABSTRACT:

Objective: To evaluate the reproducibility of the EFI (Endometriosis Fertility Index).

Design: Single-cohort prospective observational study.

Setting: University hospital.

Population: Women undergoing laparoscopic resection of any rASRM-stage endometriosis.

Methods: Details of pre- and per-operative findings were collected into a coded research file. EFI-scoring was performed 'en-bloc' by three different raters (expert-1 (C.T.), expert-2 (C.M.), junior (C.B.)). Required sample size: 71. Definitions used for agreement: clinical (scores within same range: 0-4, 5-6, 7-10) and numerical (difference ≤ 1 EFI-point).

Main outcome measures: Primary outcome: rate of clinical agreement between two experts.

Secondary outcomes: expert numerical agreement, clinical and numerical agreement between expert-1 and junior and within expert-1 (intra-observer), agreement of rASRM-score and -stage.

Results: A near-to-perfect 'inter-expert' clinical agreement rate (1.000 (95% CI 0.956-1.000), p=0.0149) was observed. The numerical agreement between two experts was also high (0.988 (95% CI 0.934-1.000)); similarly high agreement rates were observed for both 'junior-expert' comparison (clinical 0 .963 (95% CI 0.897-0.992), numerical 0.988 (95% CI 0.934-1.000) and 'intra-expert' comparisons (clinical 0.988 (95% CI 0.934-1.000); numerical 1.000 (95% CI 0.956-1.000)). Reasons for disagreements were different scoring of the least-function score and disagreements in rASRM-scores. The reproducibility of the rASRM-score was clearly inferior to that of the EFI for all comparisons.

Conclusion: The EFI can be reproduced reliably by different raters, further supporting its use in daily clinical practice as the principal clinical tool for postoperative fertility counselling/management of women with endometriosis.

Funding: FWO (Fund for Scientific Research, Flanders)

Keywords: Endometriosis, laparoscopy, infertility, classification, reproducibility of results

TWEETABLE ABSTRACT:

A study confirming the high reproducibility of the EFI substantiates its use in daily clinical practice.

INTRODUCTION

Although the rASRM (revised American Society for Reproductive Medicine) score[1] is the most frequently used surgical staging system for endometriosis to date, it has some serious limitations. First, its reproducibility has only been described as being 'fair to good'[2-5], thus prone to inter-observer variability. Second, it is not effective for predicting clinical outcomes of treatment, especially pregnancy rates in infertile patients.[6-8] For the latter reason, in 2010 Adamson and Pasta developed the EFI (Endometriosis Fertility Index), which now is a thoroughly validated scoring system that predicts pregnancy rates without using ART (assisted reproductive technology) treatment in postoperative endometriosis patients who suffer from infertility and takes into account all endometriosis rASRM stages.[9-13] Consequently, the EFI has been adopted by the WES (World Endometriosis Society) in their consensus on the classification of endometriosis.[14] In the EFI, 5 out of 10 possible points are based on patient characteristics such as age, duration of infertility and history of pregnancy. Parts of the rASRM staging account for 2 points of the EFI. Being an end-of-surgery staging, the rest of the score is based on visual observation and qualitative assessment by the surgeon (adnexal 'least function' score: 3 points). Especially the surgical part of the EFI score could make it prone to differences in interpretations by different observers, which in turn could have an effect on subsequent patient management. In the paper by Adamson and Pasta[9] who developed the EFI, a sensitivity analysis was reported to assess the effect on the EFI of potentially assumed differences in the assignment of the adnexal least function score by different surgeons, it was concluded that an EFI change of more than 1 point would only be present in 5.4% of the cases; the authors further stated that changes in the EFI would be material only for the middle values. However, this was only a theoretical exercise, and a possible added influence of the poor inter-observer agreement of the rASRM score and stage was not accounted for. Also, to our knowledge, no true inter-observer variability/reliability assessment for the EFI has been performed so far.

The objective of this study was to evaluate whether the EFI score can be reproduced reliably by different raters, i.e. whether the inter-observer variability is absent or low enough to avoid a relevant impact on clinical patient management. Additionally, intra-observer agreement of the EFI, and inter- and intra-observer agreement on the rASRM score were also studied.

METHODS

Study design

This is a single cohort prospective observational (non-interventional) study in women scheduled for endometriosis surgery of any rASRM stage at the LUFC (Leuven University Fertility Centre) of the University Hospitals Leuven Belgium). The study was conducted, based on patient data gathered from surgical procedures performed from June 13th, 2016 until December 22nd, 2016 included. Three assessors with a different profile were chosen: C.T. is an expert surgeon with a long experience of EFI-scoring, C.M. is also an expert surgeon who only occasionally uses the EFI score, and C.B. is a trainee in obstetrics and gynaecology.

Three different comparison levels were decided when designing the study protocol: comparison between rating of expert 1 (C.T.) and expert 2 (C.M.) (further referred to as 'inter-expert'), between rating of expert 1 (C.T.) and junior (C.B.) ('junior-expert'), and between rating of the first and the second session of expert 1 (C.T.) ('intra-expert').

The choice of experts as well as a trainee makes this study interesting not only for a tertiary referral centre for endometriosis, but also for those with less experience with the disease (such as trainees). There was no involvement from patients or public in the development of this study.

Study population – eligibility criteria

The LUFC is a tertiary referral centre for both endometriosis and reproductive medicine. Women of the reproductive age group (18-45 years), undergoing CO_2-laser laparoscopic surgery at the LUFC for diagnosis and treatment of endometriosis, with confirmed diagnosis on pathological examination, were eligible for this study. Indication for surgery had to be at least one of the following: infertility of ≥12 months, clinical examination and/or pain symptoms suggesting endometriosis, ultrasound (and/or other relevant imaging) findings suggesting endometriosis, previous surgical diagnosis of endometriosis. Laparoscopic procedures in the setting of a day surgery centre as well as a hospitalization setting were included. Patients were excluded in case they had a history of or were planned to undergo a hysterectomy and/or bilateral salpingo-oophorectomy, if endometriosis lesions were not completely resected (e.g. only marsupialization of an endometrioma), if photographic documentation was not performed or not compatible with study quality standards (see description in study procedures), or if informed consent was not obtained.

No extra study-related patient informed consent was necessary, since patients agreed preoperatively in their surgical informed consent form that their clinical data (which routinely include photographic documentation of the surgery) may be stored and used for scientific purposes. Confidentiality was ascertained by anonymously transferring the necessary patient data into a specifically designed research file (CRF).

Data recording and procedures

Next to demographical and clinical data (including results from clinical examination, imaging, extensive surgical reports and those specific data necessary for calculation of the historical part of the EFI), standardised photographic documentation of the laparoscopic findings was done, both at the start and end of the surgery as per WERF-EPHect-guidelines.[15] Although no video recordings were used, the mobility of the tube and ovary was be assessed on photograph by lifting the adnexa out of the ovarian fossa.

All necessary data were transferred to the CRF by C.B., a second-year obstetrics and gynaecology resident-in-training at the time of the study. In this CRF, data were anonymized and standardized, information on date of surgery was removed, and a unique and anonymous study number was allocated to each patient, to guarantee confidentiality and blinding of the assessors.

Surgical procedures were performed by C.T. or C.M., both reproductive endocrinologists as well as reproductive gynaecological surgeons with a specific expertise in the treatment of all forms of endometriosis. [16]

Only when the appropriate sample size was reached and subsequently all CRFs had been created, 'en-bloc' rating sessions were organized for each rater. All raters scored the EFI based on all the information in the CRF separately and independently from each other. Completed scoring forms were kept under lock by the study coordinator until the time of data analysis. There was at least four weeks between the last surgical procedure and the first rating session. Recall or other bias of the raters was avoided due to the time interval between surgery and rating session, the anonymization of the patient information in the CRF, the different order in which the files were rated, and the closed storage of the completed scoring forms.

During the rating session, all raters completed two scoring forms per patient: one for the rASRM and one for the surgical part of the EFI, based on the pre- and per-operative information in the CRF. Four weeks after her first rating, C.T. repeated the rating session for intra-observer variability assessment. Since the historical EFI factors are not prone to be interpreted differently by different observers, they were filled directly into the final study database but weren't scored by each rater separately. For the final calculation of the total

EFI score for each patient and for each rater/session, the (fixed) historical and (differentially rated) surgical EFI points were added together in the study database.

Outcomes

The primary outcome studied was the percentage of clinical agreement of the EFI-score in the 'inter-expert' comparison. Clinical agreement was defined as having no impact on the subsequent clinical decision pathway regarding fertility management as currently used at the LUFC, meaning that EFI-scores should be within the same range (low EFI range: 0-4, median EFI range: 5-6, high EFI range: 7-10).

Secondary outcomes studied were: clinical agreement on the EFI-score for 'junior- expert' and 'intra-expert' comparison, numerical agreement on the EFI-score (defined as a maximally allowed absolute difference in EFI-score of 1 point, regardless of the above mentioned range) and agreement on rASRM-score/stage for all three comparisons ('inter-expert', 'junior-expert', 'intra-expert').

Sample size estimation

This study was designed to show that the percentage of agreement between two senior raters (inter-expert comparison) is higher than 95% for clinical agreement (primary outcome). Based on a one-sided binomial test for a single proportion with alpha=0.05, expecting the true percentage of discrepancies to be <0.001%, the minimal sample size equals 71 subjects to have at least 80% power to show that the percentage of discrepancies is lower than 5%. The minimally required sample size was therefore set at 71.

Statistical analysis

A one-sided binomial test with alpha=0.05 for a single proportion was used to test if the observed proportion of clinical agreement between both experts was significantly higher than 95%. For all percentages of agreement, two-sided 95% CIs are reported as well. Weighted kappas (with the classical quadratic weighing), which are widely used in agreement studies[17-21], were reported both for the total EFI and for the rASRM stage, where a kappa of 1 indicates perfect agreement and 0 indicates agreement equivalent to chance. Bland-Altman plots were used to visualize the agreement of the total rASRM score. [22] Such plots provide information on the bias (the mean difference as tested with a paired t-test), the expected range of the difference in scores (95% LOA (limits of agreement)) and the possible dependency of the difference on the level of the score. Additionally, the ICC (intra-class correlation coefficient) was given for the quantification of the agreement for the total rASRM score.[23]

All analyses have been performed using SAS software, version 9.4 of the SAS System for Windows.

RESULTS

156 patients underwent laparoscopic surgery at the LUFC between June 13th, 2016 until December 22nd, 2016 included. 29 patients did not have endometriosis at laparoscopy. Out of 127 laparoscopies for endometriosis, 10 did not fit the inclusion criteria: 2 patients were outside age range, 3 had incomplete surgery for the pelvis, 4 underwent planned 2-step surgery and 1 patient had additional pathology. Out of the 117 eligible patients, 35 did not have sufficiently detailed photographic documentation, so finally 82 patients were included for creation of CRFs, rating and analysis, which was more than the minimally required

sample size. Among the included patients, 41 surgical procedures were performed by C.T., and 41 by C.M.; 13 were assisted by C.B..

EFI

Baseline demographic characteristics, including those necessary for calculation of the historical points of the EFI, are shown in Table 1. The most frequently found type of endometriosis lesions were peritoneal implants (78/82, 90.2%), followed by deep (64/82, 78.1%), superficial ovarian 42/82 (51.2%) and cystic ovarian (23/82, 28%).

Table 2A shows the results for EFI agreement according to both definitions described above, and the weighted kappa for the 3 comparisons made. The majority of included patients had high scores for the historical part of the EFI (4 points, 45/82 (54.88%) or 5 points 23/82 (28.05%)), as reflected partly in the clustering of the higher EFI-scores (Table 3). This is comparable with a previous study in our population [10], which confirms the studied population as representative for our clinic.

Inter-expert EFI comparison

For the ‘inter-expert’ clinical agreement, the study hypothesis was confirmed, namely that the rate of agreement was higher than 95%, which was near-to-perfect (1.000 (95% CI 0.956-1.000), one-sided p-value=0.0149).

The ‘inter-expert’ numerical agreement was slightly lower than the clinical agreement (with the lower limit of the 95% CI just below 0.95: 0.988 (95% CI 0.934-1.000)).

Table 3 shows the details of agreement for the 'inter-expert' comparison (similar data on the other comparisons can be supplied upon request). In 9 cases, EFI scores did not reach absolute agreement between both experts C.T. and C.M., of which only 1 led to the defined 'numerical disagreement' (EFI score 4 versus 2). Out of these 9 cases, 3 were due to differences in rASRM score (1 in lesion score <or≥16, 2 in total score <or≥71), and 6 were due to C.T. giving a lower LF score than C.M. (4 with bilateral vaporization of superficial ovarian endometriosis, 1 with treatment of an endometrioma, and 1 for of tubal/fimbrial functionality).

Junior-Expert EFI comparison

For the comparison 'junior-expert', in general the rate of agreement was slightly lower than for the inter-expert EFI comparison, but still around 90% or more when taking into account the lower limit of the 95% CI (0.963 (95% CI 0.897-0.992) for clinical agreement, 0.988 (95% CI 0.934-1.000) for numerical agreement).

Details of disagreement were as follows: 1 case with both numerical and clinical disagreement and 2 cases with clinical disagreement only, out of the total of 15/82 files with any difference in EFI scoring between junior and expert. Of these 15 cases, 4 were due to a difference in total rASRM score (> or ≤71), 7 due to different ovarian LF score (of which 1 led to clinical disagreement) and 4 due to different tubal/fimbrial LF score (of which 1 led to clinical, and 1 to clinical and numerical disagreement).

Intra-expert EFI comparison

Agreement was also high for the ‘intra-expert’ comparison (numerical agreement (1.000 (95% CI 0.956-1.000), clinical agreement (0.988 (95% CI 0.934-1.000)).

For this comparison, only 1 case had clinical disagreement out of a total of 7/82 of cases with any difference in EFI score. Of these latter 7 cases, 1 difference was attributed to the total rASRM score, and 6 to the LF score (4 on ovarian function and 2 on tubal/fimbrial function (amongst which 1 led to clinical disagreement)).

rASRM scoring and staging

From Figure 1, showing the Bland-Altman plot and statistical analysis of the agreement on the total rASRM score (in points), it’s clear that the variability for the total rASRM score given is very large for all 3 comparisons. Indeed, although the mean differences of assigned rASRM points may be small (confirming a low risk for fixed bias), their SDs are large, and the 95% LOA (limits of agreement) span a width of 40 points or more, which is comparable to 4 rASRM stages.

Table 2B describes the analysis of agreement on rASRM stage, explained by rate of agreement and weighted kappa; these results are consistently lower than those obtained for the EFI (Table 2A). Figure S1 shows an example of a woman where complete agreement between all raters was found.

Relationship between rASRM and EFI

Figure S2 shows a boxplot of the distribution of rASRM total score for each EFI range (for expert 1). This illustrates that in general there was a negative correlation between the rASRM (points/stage) and EFI range. Interestingly, 43/62 (72,58%) of women with a high EFI also have rASRM stage III-IV endometriosis (Figure S2B).

DISCUSSION

Main findings

Our study represents the first report on inter- and intra-observer reproducibility of the EFI and demonstrates high intra- and inter-agreement rate with narrow 95%CIs. More specifically, we confirmed our hypothesis that clinical agreement for the 'inter-expert' comparison (primary outcome) was higher than 95%. These results concur with the hypothetical assumption based on the sensitivity analysis on the EFI by Adamson and Pasta[9]. In addition, very high agreements were also reported for numerical 'inter-expert' agreement, clinical and numerical 'junior-expert' and 'intra-expert' comparison (secondary outcomes), although not near-to-perfect as for clinical "inter-expert" agreement. In other words, the high reproducibility supports the use of the EFI in daily clinical practice as a very relevant clinical tool for management and counselling of postoperative endometriosis patients on their reproductive outcome.

Strengths and limitations

Our study was designed to avoid bias in several ways. First of all, the assessment of the EFI was done based on a combination of patient history information, standardized operative reports and complete photographic series of the operative site, in order to prevent any misclassification of rASRM staging and associated adnexal adhesions as much as possible.[5, 21] Second, to blind raters to the personal details of patients, a coded CRF was used for rating instead of the patient file itself. Third, to avoid recall bias, a standardized and anonymized CRF was used. Additionally, 'en-bloc' rating sessions, with random order of patient files, were organised for each rater. Fourth, since C.T. had the most experience in calculating the EFI in clinical practice, her first rating was therefore chosen as standard to assess agreement with the second expert ('inter-expert'), the junior surgeon ('junior-expert') and within one rater ('intra-expert').

Out of the 117 eligible patients, 35 were excluded because they did not have sufficiently detailed photographic documentation. This was not considered as a flaw, but merely a consequence of the fact that the study was conducted in a real life turbulent clinical setting (different surgeons, different operation theatres, technical difficulties etc.). Patients files were only included if photographic documentation (both pre- and postoperative) met the criteria as defined per WERF-EpHect procedures.[15] Despite this strict selection, our study population was still representative for the population in our clinic (see result section), and the minimally required sample size was more than met.

This study has a number of limitations that should be taken into account. First, the relatively small numbers of raters involved may be a negative point, although this was accounted for in the sample size calculation as discussed in the methodology section. Second, raters with various levels of expertise of EFI scoring were included as describe in the methodology section. The junior rater was also trained by the expert rater amongst others. Therefore, we would suggest future studies on the reproducibility of the EFI to include a larger number of observers and a more varied pool of observers preferably from various centres with different expertise. Third, risk of recall bias cannot completely be excluded since both experts performed all laparoscopies, and the junior assisted some procedures. Fourth, the use of photographic documentation only rather than video recording during the surgery to assess both the initial endometriosis lesions and the least function score at conclusion of surgery may be less precise. However, as per WERF-EPHect-guidelines[15], photographic documentation only was assumed to be sufficient for the aim of our study and could easily be embedded in our daily clinical practice. Fifth, next to photographic documentation, standardized operative reports were provided to the raters, which could positively influence the precision of the rating as described in the inter-rater agreement study of Schliep et al[21]. However, this argument can easily be rejected since – in contrast to the EFI – the reproducibility of rASRM score and stage remained poor. Finally, the estimated sample size for the primary outcome (i.e. percentage of clinical agreement) may appear too small, although the null hypothesis was derived from the EFI development study[9]. In hindsight, the assumption used in the calculation (true percentage of discrepancies lower than 0.001%) could be considered as too optimistic.

Interpretation

Disagreement between raters could be largely explained by differential rating of the least function score, and of the rASRM score. The influence of lower reproducibility of the rASRM score on the EFI score reproducibility was not taken into account in the sensitivity analysis by Adamson and Pasta[9] but is now identified in our data – next to the least function score – as a potential weak spot in the reproducibility of the EFI score.

For all comparisons made, the rate of agreement was lower for the rASRM endometriosis total score and rASRM endometriosis stage than for EFI score, despite our efforts to avoid misclassification as described above. With respect to assessments of rASRM total score, the width of variation was very high, and therefore the finding of a low mean error for all three comparisons is not necessarily reassuring. Indeed, also ICCs are falsely inflated, since they compare the difference within a subject to the difference between subjects, and in a more uniform population (where the range of rASRM total score would be smaller than in our population) the ICC would be considerably lower if still similar variation between observers would be found.

With respect to rASRM stage assessment, agreements were also lower than for the EFI, as explained by the lower values for weighted kappa and the lower limits of 95% CI for agreement per se. When comparing results for weighted kappa, it should be noted that, in contrast to the EFI where 11 possible categories are withheld (0-10, including both), in the rASRM classification only 4 stages are categorized, but still results on rASRM stage showed a markedly higher variability.

CONCLUSION

In addition to already vast evidence confirming the EFI score to be superior to the rASRM score/stage for the prediction of reproductive outcome after surgery, our study has now clearly demonstrated that EFI scoring is highly reproducible. This high reproducibility is far better than for the rASRM scoring/staging, even for a trainee. Collectively, this evidence supports the standard use of the EFI score next to the rASRM score/stage in daily clinical practice as also advised by the WES [14], and the replacement of the rASRM stage/score by the EFI score for postoperative fertility counselling of endometriosis patients. Preferably, our data on reproducibility of the EFI score, as presented in this study, should be confirmed by other groups, ideally by using a similar methodology but with a larger number of raters to enhance comparability with our data.

ACKNOWLEDGEMENTS

None

DISCLOSURE OF INTERESTS

- The LUFC received unrestricted research grants from Ferring Pharmaceuticals and Merck SA.
- C.T. received consultancy fees from Gedeon Richter, Lumenis and Nordic Pharma (payed to UZ/KU Leuven, no private revenue). Sponsoring from Ferring Pharmaceuticals, Merck SA, Gedeon Richter and Bayer to travel to and attend scientific meetings.
- C.B., M.W. and S.F. report no conflicts of interest
- C.M. received consultancy fees from Lumenis (payed to KULeuven, no private revenue)
- T.D. has become vice-president and head of global medical affairs infertility for the multinational pharmaceutical company Merck (Darmstadt, Germany) from October 1st, 2015. He continues on a part time basis his academic appointment as Professor of Reproductive Medicine at the University of Leuven (KU Leuven) in Belgium and as Adjunct Professor at the Department of Obstetrics and Gynaecology at Yale University, New Haven, CO, USA.

- Completed disclosure of interest forms are available to view online as supporting information.

CONTRIBUTION TO AUTHORSHIP

C.T.: study design, study performance (surgery, CRF design, rating), article writing and editing

C.B.: study performance (surgery, CRF data-transfer, rating), article editing

C.M.: study performance (surgery, rating), article editing

M.W.: study performance (CRF data-transfer, file coding), article editing

S.F.: statistical analysis, article editing

T.D.: study design, article editing

DETAILS OF ETHICS APPROVAL

This study was approved by the ethics committee of the UZ Leuven on June 8th, 2016 (internal UZ Leuven Trial Registration Number: S59221); competent authority approval was not necessary since the study was observational.

FUNDING

This study was supported by FWO (Fund for Scientific Research, Flanders).

REFERENCES

1. American Society for Reproductive M. Revised American Society for Reproductive Medicine classification of endometriosis: 1996. Fertility and Sterility. 1997;67(5):817-21.

2. Rock JA, Zoladex Endometriosis Study Group JA. The revised American Fertility Society classification of endometriosis: reproducibility of scoring. Fertility and Sterility. 1995;63(5):1108-10.

3. 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(5):1015-21.

4. Lin SY, Lee RKK, Hwu YM, Lin MH. Reproducibility of the revised American Fertility Society classification of endometriosis using laparoscopy or laparotomy. International Journal of Gynecology & Obstetrics. 1998;60(3):265-9.

5. Schliep CK, Stanford BJ, Chen KZ, Zhang OB, Dorais WJ, Boiman Johnstone BE, et al. Interrater and Intrarater Reliability in the Diagnosis and Staging of Endometriosis. Obstetrics & Gynecology. 2012;120(1):104-12.

6. Palmisano GP, Adamson GD, Lamb EJ. Can staging systems for endometriosis based on anatomic location and lesion type predict pregnancy rates? Int J Fertil Menopausal Stud. 1993;38(4):241-9.

7. Vercellini P, Fedele L, Aimi G, De Giorgi O, Consonni D, Crosignani PG. Reproductive performance, pain recurrence and disease relapse after conservative surgical treatment for endometriosis: the predictive value of the current classification system. Human Reproduction. 2006;21(10):2679-85.

8. Adamson DG. Endometriosis classification: an update. Current Opinion in Obstetrics and Gynecology. 2011;23(4):213-20.

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

10. Tomassetti C, Geysenbergh B, Meuleman C, Timmerman D, Fieuws S, D'Hooghe T. External validation of the endometriosis fertility index (EFI) staging system for predicting non-ART pregnancy after endometriosis surgery. Human Reproduction. 2013;28(5):1280-8.

11. Zeng C, Xu J-N, Zhou Y, Zhou Y-F, Zhu S-N, Xue Q. Reproductive Performance after Surgery for Endometriosis: Predictive Value of the Revised American Fertility Society Classification and the Endometriosis Fertility Index. Gynecologic and Obstetric Investigation. 2014;77(3):180-5.

12. Garavaglia E, Pagliardini L, Tandoi I, Sigismondi C, Viganò P, Ferrari S, et al. External Validation of the Endometriosis Fertility Index (EFI) for Predicting Spontaneous Pregnancy after Surgery: Further Considerations on Its Validity. Gynecologic and Obstetric Investigation. 2015;79(2):113-8.

13. Boujenah J, Bonneau C, Hugues JN, Sifer C, Poncelet C. External validation of the Endometriosis Fertility Index in a French population. Fertility and Sterility. 2015;104(1):119-23.e1.

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

15. Becker CM, Laufer MR, Stratton P, Hummelshoj L, Missmer SA, Zondervan KT, et al. World Endometriosis Research Foundation Endometriosis Phenome and Biobanking Harmonisation Project: I. Surgical phenotype data collection in endometriosis research. Fertility and Sterility. 2014;102(5):1213-22.

16. Meuleman C, Tomassetti C, Wolthuis A, Van Cleynenbreugel B, Laenen A, Penninckx F, et al. Clinical outcome after radical excision of moderate-severe endometriosis with or without bowel resection and reanastomosis: A prospective cohort study. Annals of Surgery. 2014;259(3):522-31.

17. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005;37(5):360-3.

18. Holland TK, Hoo WL, Mavrelos D, Saridogan E, Cutner A, Jurkovic D. Reproducibility of assessment of severity of pelvic endometriosis using transvaginal ultrasound. Ultrasound Obstet Gynecol. 2013;41:210-5.

19. Zahn CM, Luigi KFR, Olsen C, Whitworth SA, Washingtom A, Crothers B. Reproducibility of Endocervical Curettage Diagnosis. Obstet Gynecol. 2011;118:240-8.

20. Paternot G, Devroe J, Debrock S, D'Hooghe TM, Spiessens C. Intra- and inter-observer analysis in the morphological assessment of early-stage embryos. Reproductive Biology and Endocrinology. 2009;7:105.

21. Schliep K, Chen Z, Stanford J, Xie Y, Mumford S, Hammoud A, et al. Endometriosis diagnosis and staging by operating surgeon and expert review using multiple diagnostic tools: an inter-rater agreement study. BJOG: An International Journal of Obstetrics & Gynaecology. 2017;124(2):220-9.

22. Bland JM, Altman DG. Measuring agreement in method comparison studies. Statistical Methods in Medical Research. 1999;8(2):135-60.

23. McGraw KO, Wong SP. Forming Inferences About Some Intraclass Correlation Coefficients. Psychological Methods. 1996;1(1):30-46.

Table 1. Baseline characteristics, including historical factors of the EFI and their translation into EFI-points, for the total population (N=82) (NA = not applicable)

Characteristic	Mean ± SD	Median (IQR)	Number of patients/total (%)
Pain symptoms	NA	NA	
- Dysmenorrhea			75/82 (91,5%)
- Dyschezia			45/82 (54,9%)
- Rectal bleeding			16/82 (19,5%)
- Deep dyspareunia			37/81 (45,7%)
- Chronic pelvic pain			36/82 (43,9%)
- Mictalgia			24/82 (29,3%)
History of diagnostic/incomplete surgery	NA	NA	39/82 (47,5%)
History of fertility treatment	NA	NA	
- IUI			15/82 (18,29%)
- ART			13/82 (15,85%)
Age (in years)	31.5 ± 4.65	31.2 (28.4-34.8)	0 EFI points (age 40+): 1/82(1.22%) 1 EFI point (age 36-39): 16/82 (19.51%) 2 EFI points (age <36): 65/82 (79.27%)
Duration of infertility (in months)	17.1 ± 22.17	13.0 (0-29)	0 EFI points (>3 years): 7/82 (8.54%) 1 EFI point (≤3 years): 75/82 (91.46%)
Prior pregnancy	NA	NA	0 EFI point (never): 49/82 (59.76%) 1 EFI point (ever): 33/82 (40.24%)
EFI: total historical points	NA	NA	0 EFI points: 0/82 (0%) 1 EFI point: 1/82 (1.2%) 2 EFI points: 6/82 (7.3%) 3 EFI points: 7/82 (8.5%) 4 EFI points:45/82 (54.5%) 5 EFI points: 23/82 (28.1%)

Table 2. Agreement for total EFI score and rASRM stage between raters

Table 2A: Agreement for total EFI score between raters

Comparison	**Clinical agreement** **EFI score** **(95% CI)**	**Numerical agreement** **EFI score** **(95% CI)**	**Weighted kappa** **EFI score** **(95% CI)**
Inter-expert	**1.000 (0.956-1.000) ***	0.988 (0.934-1.000)	0.942 (0.904-0.980)
Junior-expert	0.963 (0.897-0.992)	0.988 (0.934-1.000)	0.907 (0.858-0.956)
Intra-expert	0.988 (0.934-1.000)	1.000 (0.956-1.000)	0.959 (0.929-0.990)

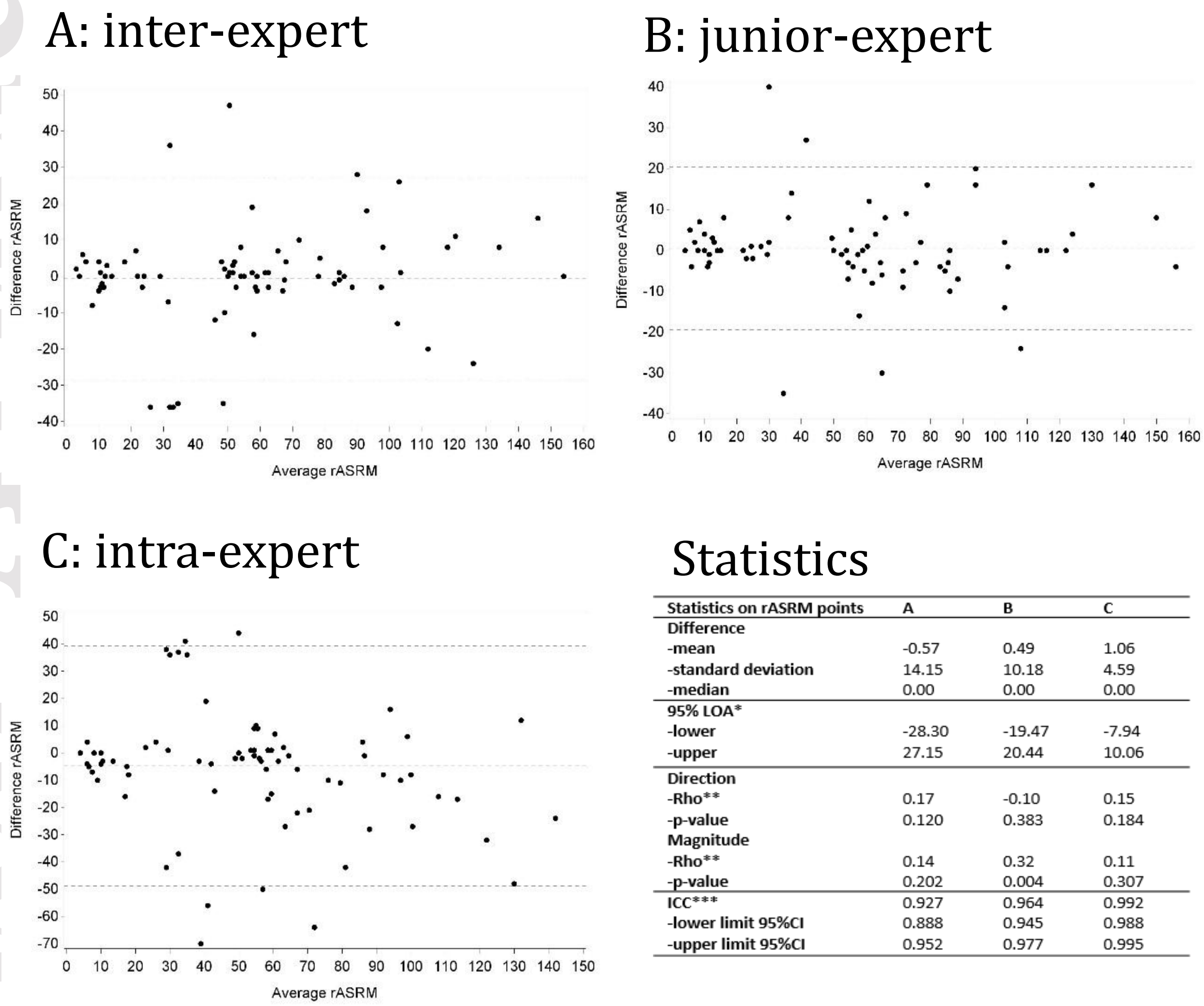
*primary outcome: one-sided p-value = 0.0149

Table 2B: Analysis of (absolute) agreement on rASRM stage for the different comparisons

Comparison	**Agreement** **rASRM stage** **(95% CI)**	**Weighted kappa** **rASRM stage** **(95% CI)**
Inter-expert	0.841 (0.744-0.913)	0.752 (0.621-0.882)
Junior-expert	0.890 (0.802-0.949)	0.752 (0.721-0.882)
Intra-expert	0.915 (0.832-0.965)	0.907 (0.847-0.968)

Table 3. Cross-tabulation of the frequency of a given EFI-score for the inter-expert comparison – raw data (note that no score below 2 was given by any of the two raters).

		EFI by expert 2									
		2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
EFI by expert 1	**2**	**1**	0	1	0	0	0	0	0	0	**2**
	3	0	**2**	0	0	0	0	0	0	0	**2**
	4	0	0	**3**	0	0	0	0	0	0	**3**
	5	0	0	0	**7**	0	0	0	0	0	**7**
	6	0	0	0	0	**6**	0	0	0	0	**6**
	7	0	0	0	0	0	**12**	2	0	0	**14**
	8	0	0	0	0	0	2	**13**	4	0	**19**
	9	0	0	0	0	0	0	0	**19**	0	**19**
	10	0	0	0	0	0	0	0	0	**10**	**10**
	Total	1	2	4	7	6	14	15	23	10	**82**



Statistics on rASRM points	A	B	C
Difference			
-mean	-0.57	0.49	1.06
-standard deviation	14.15	10.18	4.59
-median	0.00	0.00	0.00
95% LOA*			
-lower	-28.30	-19.47	-7.94
-upper	27.15	20.44	10.06
Direction			
-Rho**	0.17	-0.10	0.15
-p-value	0.120	0.383	0.184
Magnitude			
-Rho**	0.14	0.32	0.11
-p-value	0.202	0.004	0.307
ICC***	0.927	0.964	0.992
-lower limit 95%CI	0.888	0.945	0.988
-upper limit 95%CI	0.952	0.977	0.995

Figure 1. Agreement in total rASRM score per comparison: Bland-Altman plots and statistics.

*LOA = limits of agreement;

**Rho = spearman correlation coefficient

***ICC = single measure intra-class correlation coefficient, 2-way random/mixed model with absolute agreement definition