RESEARCH

Deep learning to diagnose pouch of Douglas obliteration with ultrasound sliding sign

Gabriel Maicas[1],*, Mathew Leonardi[2,3,4],*, Jodie Avery[5], Catrina Panuccio[6], Gustavo Carneiro[1], M Louise Hull[5,7] and George Condous[2,3]

[1]Australian Institute for Machine Learning, University of Adelaide, Adelaide, Australia
[2]OMNI Gynaecological Ultrasound and Care, Sydney, Australia
[3]Sydney Medical School Nepean, University of Sydney, Sydney, Australia
[4]Department of Obstetrics and Gynecology, McMaster University, Hamilton, Canada
[5]Robinson Research Institute, University of Adelaide, Adelaide, Australia
[6]Specialist Imaging Partners, North Adelaide, Australia
[7]Discipline of Obstetrics and Gynaecology, Women & Children's Hospital, Adelaide, Australia

Correspondence should be addressed to M Leonardi: **Mathew.Leonardi@sydney.edu.au**

*(G Maicas and M Leonardi contributed equally to this paper and should be regarded as joint first authors)

Abstract

Objectives: Pouch of Douglas (POD) obliteration is a severe consequence of inflammation in the pelvis, often seen in patients with endometriosis. The sliding sign is a dynamic transvaginal ultrasound (TVS) test that can diagnose POD obliteration. We aimed to develop a deep learning (DL) model to automatically classify the state of the POD using recorded videos depicting the sliding sign test.
Methods: Two expert sonologists performed, interpreted, and recorded videos of consecutive patients from September 2018 to April 2020. The sliding sign was classified as positive (i.e. normal) or negative (i.e. abnormal; POD obliteration). A DL model based on a temporal residual network was prospectively trained with a dataset of TVS videos. The model was tested on an independent test set and its diagnostic accuracy including area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, and positive and negative predictive value (PPV/NPV) was compared to the reference standard sonologist classification (positive or negative sliding sign).
Results: In a dataset consisting of 749 videos, a positive sliding sign was depicted in 646 (86.2%) videos, whereas 103 (13.8%) videos depicted a negative sliding sign. The dataset was split into training (414 videos), validation (139), and testing (196) maintaining similar positive/negative proportions. When applied to the test dataset using a threshold of 0.9, the model achieved: AUC 96.5% (95% CI: 90.8–100.0%), an accuracy of 88.8% (95% CI: 83.5–92.8%), sensitivity of 88.6% (95% CI: 83.0–92.9%), specificity of 90.0% (95% CI: 68.3–98.8%), a PPV of 98.7% (95% CI: 95.4–99.7%), and an NPV of 47.7% (95% CI: 36.8–58.2%).
Conclusions: We have developed an accurate DL model for the prediction of the TVS-based sliding sign classification.

Lay summary

Endometriosis is a disease that affects females. It can cause very severe scarring inside the body, especially in the pelvis – called the pouch of Douglas (POD). An ultrasound test called the 'sliding sign' can diagnose POD scarring. In our study, we provided input to a computer on how to interpret the sliding sign and determine whether there was POD scarring or not. This is a type of artificial intelligence called deep learning (DL). For this purpose, two expert ultrasound specialists recorded 749 videos of the sliding sign. Most of them (646) were normal and 103 showed POD scarring. In order for the computer to interpret, both normal and abnormal videos were required. After providing the necessary inputs to the computer, the DL

model was very accurate (almost nine out of every ten videos was correctly determined by the DL model). In conclusion, we have developed an artificial intelligence that can interpret ultrasound videos of the sliding sign that show POD scarring that is almost as accurate as the ultrasound specialists. We believe this could help increase the knowledge on POD scarring in people with endometriosis.

Introduction

The pouch of Douglas (POD) is a space in the female pelvis between the retrocervix and the anterior rectum and between the uterosacral ligaments. The space may be obliterated by adhesions, usually including the uterus and rectum, leading to an inability to visualize the peritoneum (Cullen 1914). Obliteration exists in several scenarios: endometriosis, infections, malignancy, and iatrogenic surgical adhesions. Research on POD obliteration usually focuses on endometriosis due to its role in disease stage classification and surgical implications, such as incomplete surgery resulting in residual disease or intraoperative complications (Melnyk *et al.* 2020, Espada *et al.* 2021). Nonetheless, POD obliteration is a pertinent state to be aware of pre-operatively for all pelvic surgery as it increases the surgical complexity and is associated with complications (Purohit *et al.* 2018, Leonardi *et al.* 2020a).

The sliding sign is an accurate dynamic transvaginal ultrasound (TVS) test that is used to evaluate the POD (Hudelist *et al.* 2013, Reid *et al.* 2013). It can be interpreted by an ultrasound operator at the time of point-of-care scanning or by a radiologist/sonologist observing the recorded videos (Chiu *et al.* 2019). The dynamic nature of TVS mandates that in order to perform the sliding sign correctly, one must have adequate knowledge of normal (producing a positive sliding sign) and abnormal (producing a negative sliding sign) female pelvic anatomy.

Ultrasound has undoubtedly become indispensable in the diagnostic workup of gynecologic pathology, including endometriosis (Nisenblat *et al.* 2016), but several flaws with this imaging modality exist. Most notably, it relies on an operator and diagnostician expertise, along with which comes variable inter and intraobserver accuracy (Menakaya *et al.* 2016). Expertise becomes even more relevant as these new techniques, which are not yet widely adopted (Leonardi *et al.* 2020c), have a learning curve to achieve competency (Tammaa *et al.* 2014, Leonardi *et al.* 2020b).

While we attempt to overcome obstacles such as interobserver variability and the learning curve to become competent with a new concept, deep learning (DL), a branch of machine learning, could be considered as a method of computer-aided classification to encourage more rapid adoption of the sliding sign technique (Drukker *et al.* 2018). The main advantage of DL methods is that the features are automatically learned to maximize the classification performance. We aimed to develop a DL model to automatically classify the state of the POD using the sliding sign test.

Materials and methods

Study design

A prospective diagnostic accuracy study was performed and reported according to the STARD guidelines (Bossuyt *et al.* 2015).

Setting

The study was performed at a high-volume gynecology-focused ultrasound practice in Sydney, Australia between September 2018 and April 2020. Equipment consisted of GE Healthcare Voluson E8 or S6 ultrasound machines (General Electric, Zipf, Austria) with 4–9 MHz transvaginal transducers. All data were recorded using GE Healthcare ViewPoint (General Electric, Zipf, Austria).

Participants

We included consecutive women of all ages visiting the clinic with any indication for gynecologic TVS during the study period. The exclusion criteria included the inability to perform a TVS, history of a hysterectomy, inability to perform the sliding sign due to large pelvic pathology limiting the adequate assessment of the POD, or pregnancy with a gestational age greater than 10 weeks. Women provided a verbal consent before undergoing a TVS. This

Reproduction
&Fertility

G Maicas, M Leonardi *et al.*

Deep learning ultrasound sliding sign

2:4

238

study was approved by the Nepean Blue Mountains Local Health District ethics committee; HREC/16/Nepean/31.

Ultrasound protocol

All TVS examinations were completed by one of two gynecologic sonologists, both of whom were considered experts in the performance and interpretation of the sliding sign. They were considered as level 2 and level 3 experts as per the European Federation of Societies for Ultrasound in Medicine and Biology (European Federation of Societies for Ultrasound in Medicine and Biology 2006), respectively, at the time of the study. The method to perform the sliding sign in this study depended on the orientation of the uterus. In patients with an anteverted uterus, the technique to produce this sliding sign involves applying pressure to the fundus of the uterus (with the operator's non-scanning hand) and/or applying pressure with the tip of the probe to the cervix (Supplementary Video 1, see section on supplementary materials given at the end of this article). In patients with an axial uterus, the technique involves applying pressure with the tip of the probe to the cervix (Supplementary Video 2). In patients with a retroverted uterus, the technique involves applying pressure with the tip of the probe against the posterior uterine fundus (Supplementary Video 3). In all uterine orientations, the operators assessed the sliding of the posterior uterine and retrocervix serosa against the contents posteriorly. De-identified videos of the sliding sign were saved and the findings were interpreted by the operator on the day of the patient's visit.

Variable outcomes

The overall classification was positive when there was sliding at both the posterior uterine fundus and retrocervix, indicating a non-obliterated or normal POD state. If the sliding sign was classified as negative at one or both locations, the overall classification was negative, indicating POD obliteration. No clinical variables were collected for this study. The decision to collect only the sliding sign as an outcome variable was made because the focus of the study was to evaluate a DL method that could analyze TVS.

Machine learning approach summary

We developed a machine learning model that analyzes TVS videos depicting the sliding sign. The model received a TVS video as input and processed it to output the probability for the presence of a negative sliding sign. In the following sections, we define the dataset and the model, we describe how to train its parameters and perform inference, and we define the experimental set-up.

Dataset

Let $\{x_i, y_i\}_{i=1...|N|}$ be a dataset containing $|N|$ TVS videos, where $x:\Omega \rightarrow R$ denotes the TVS video with $\Omega \subset R^3$ representing the video lattice, and $y \in \{0,1\}$ indicates the absence ($y=0$) or presence ($y=1$) of the sliding sign. The dataset was patient-wise split into training, validation, and testing sets.

Machine learning model

We chose the state-of-the-art model Resnet (2+1)D (Tran *et al.* 2018) that showed superior performance by splitting the spatiotemporal components of the video. It consists of 18 R(2+1) convolutional layers (Tran *et al.* 2018) where each convolution is followed by a batch normalization operation (see Fig. 1 for a diagram of the model). During the training phase, the model parameters Θ are optimized by minimizing the cross-entropy loss function:

$$L_i(\theta) = -\left(y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right),$$

where, p_i is the predicted probability for the presence of the sliding sign for the ith TVS video. During inference, the probability for the presence of the sliding sign is computed by a forward pass of the model with optimal parameters. We threshold p_i at $\tau \in [0, 1]$ to decide whether an image is classified as positive (above the threshold) or negative (below the threshold).

Experimental set-up data preparation and pre-processing

The total dataset was divided into two groups using a cut-off date (December 2019): (1) training and validation, and (2) testing. The training and validation group was randomly divided into the training set (75%) and the validation set (25%). All videos in the testing dataset depicted unique patients (i.e. no videos were depicting the same patient in the training, validation, and testing sets) and each patient only had one video. See Table 1 for a summary of the dataset.

Each video had a duration of 10 s at an approximate of 30 frames per second. During pre-processing, all videos were automatically cropped by removing the first 70 rows of each frame so that it only contained the fan. We
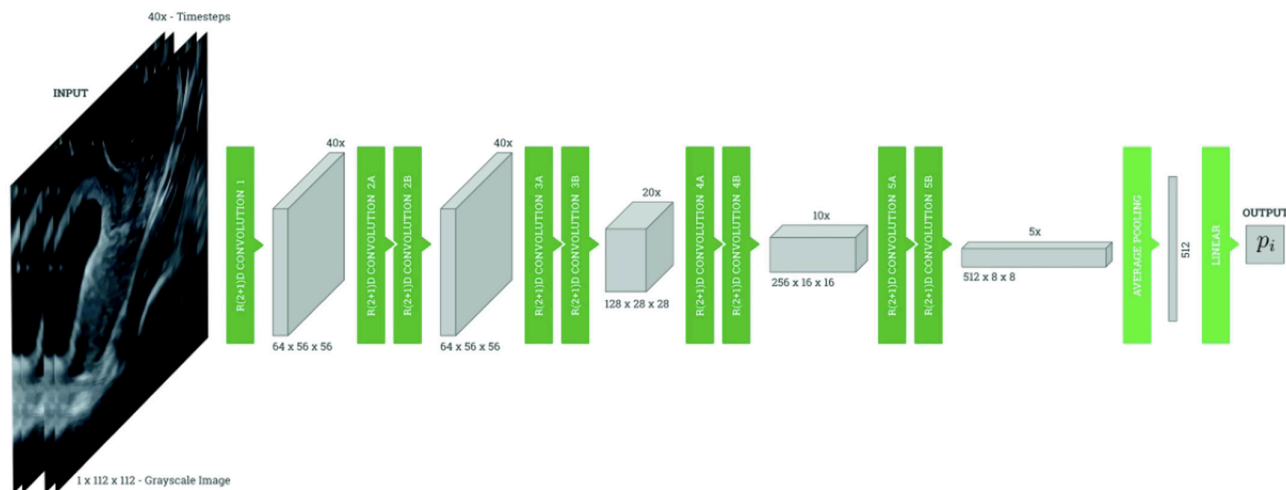
Figure 1 Graphic depiction of the deep learning (DL) model.

uniformly sampled the temporal resolution of the video to a total of 40 frames and the spatial resolution to 112 × 112 with bilinear interpolation.

The Resnet (2+1)D DL model (Tran *et al.* 2018) was pre-trained on the Kinetics-400 dataset (Paszke *et al.* 2019), and then all layers were fine-tuned on the TVS training set. Model parameters were optimized using ADAM (Kingma & Ba 2015) with a learning rate of 1e−−4 and a batch size of 5. We used the validation set for model selection, that is we chose the optimal hyperparameters for training the model based on maximizing the performance of the model in the validation set. Performance results are reported in the test set. Note that the use of a validation set is a standard practice in machine learning to tune model hyperparameters based on an unseen dataset to avoid overfitting the training data while maintaining the test set unseen during the training process. We used PyTorch (Paszke *et al.* 2019) to implement our framework.

Statistical analysis

In the test group, the diagnostic performance of DL was compared with that of the expert sonologist. Using the sonologist-apportioned sonographic classification as the

reference standard, the area under the ROC curve (AUC), accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were expressed as percentages with 95% CIs (Mercaldo *et al.* 2007, Altman *et al.* 2013). Two sets of diagnostic performance were performed to maximize the sensitivity and specificity using different thresholds τ for the p_i values as described above.

The nomenclature of the sliding sign test is opposite to the test results of most medical investigations. A normal POD is described as having a positive sliding sign and an abnormal POD (POD obliteration) is described as having a negative sliding sign. The definitions of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) are provided in the Supplementary Table 1.

A TP case is when the DL and sonologist both classify the sliding sign as positive. A TN case is when the DL and sonologist both classify the sliding sign as negative. An FN case is when the DL incorrectly classifies the sliding sign as negative but the sonologist classified it as positive. An FP case is when the DL incorrectly classifies the sliding sign as positive but the sonologist classified it as negative.

Results

Between September 2018 and April 2020, 749 sliding sign videos were recorded. The breakdown of videos in the dataset by classification (positive vs negative) is depicted in Table 1.

When applied to the test dataset, the proposed system achieved an AUC of 96.5 (95% CI: 90.8–100.0%) (Fig. 2).

Using a threshold of $\tau = 0.9$, we found an accuracy of 88.8% (95% CI: 83.5–92.8%), sensitivity of 88.6% (95% CI: 83.0–92.9%), specificity of 90.0% (95% CI: 68.3–98.8%), a

Table 1 Proportion of positive and negative sliding sign classifications in the dataset.

Datasets	*n*	Sliding sign classification, *n* (%)	
		Positive	Negative
Overall dataset	749	646 (86.2)	103 (13.8)
Training dataset	414	351 (84.8)	63 (15.2)
Validation dataset	139	119 (85.6)	20 (14.4)
Test dataset	196	176 (89.8)	20 (10.2)

PPV of 98.7% (95% CI: 95.4–99.7%), and an NPV of 47.7% (95% CI: 36.8–58.2%) (Table 2).

Using a threshold of $\tau = 0.5$, we found an accuracy 95.4% (95% CI, 91.5–97.9%), sensitivity of 98.9% (95% CI, 96.0–99.9%), specificity of 65.0% (95% CI, 40.8–84.6%), a PPV of 96.1% (95% CI, 93.2–97.8%), and an NPV of 86.7% (95% CI, 61.2–96.4%) (Table 2).

The inference time of the DL model to produce the classification of the sliding sign in a recorded video is 0.01 s using an NVIDIA Tesla K80 GPU with 24 GB of memory. The time required to transform the video to the processed resolution is 0.81 s. Thus, the total time required to perform a prediction is 0.82 s.

Discussion

Main findings

In the present study, we designed a computerized model to evaluate the sliding sign automatically in 0.82 s from TVS videos. Our proposed DL model achieved a high diagnostic performance as demonstrated by an AUC of 96.5%. Depending on the chosen threshold, the DL model can achieve various arrangements of diagnostic performance prioritizing either 'ruling in' or 'ruling out' a positive sliding sign. To avoid a false positive sliding sign when it had been deemed negative by the sonologist, we have prioritized specificity as our primary performance tool. Clinically, we feel this is more important since we do not want to miss patients with the abnormal state of POD obliteration (i.e. negative sliding sign).
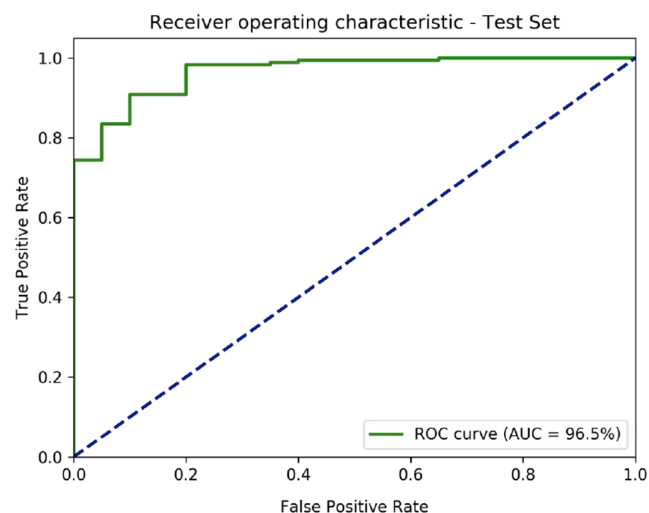

Figure 2 Receiver operating characteristic curve. ROC, receiver operating characteristic; AUC, area under the ROC curve.

Table 2 Diagnostic performance of DL to predict the classification of the sliding sign using recorded TVS videos, using thresholds of $\tau = 0.9$ and $\tau = 0.5$.

	$\tau = 0.9$	$\tau = 0.5$
True positive, n	156	174
False positive, n	2	7
True negative, n	18	13
False negative, n	20	2
Accuracy, % (95% CI)	88.8 (83.5–92.8)	95.4 (91.5–97.9)
Prevalence, % (95% CI)	89.8 (84.7–93.7)	89.8 (84.7–93.7)
Sensitivity, % (95% CI)	88.6 (83.0–92.9)	98.9 (96.0–99.9)
Specificity, % (95% CI)	90.0 (68.3–98.8)	65.0 (40.8–84.6)
PPV, % (95% CI)	98.7 (95.4–99.7)	96.1 (93.2–97.8)
NPV, % (95% CI)	47.7 (36.8–58.2)	96.7 (61.2–96.4)

NPV, negative predictive value; PPV, positive predictive value.

Interpretation

In most medical settings, the prevalence of a normal POD far outweighs the abnormal state of POD obliteration. Even in specialist endometriosis centers, the prevalence of POD obliteration ranges from 20 to 30% (Hudelist *et al.* 2013, Reid *et al.* 2013). Only recently has the importance of POD obliteration outside of endometriosis been raised; it is thought that roughly 1 in 29 women without the concern for endometriosis have POD obliteration (Leonardi *et al.* 2020*a*). It is well understood that recognizing POD obliteration non-invasively is crucial (Tompsett *et al.* 2019, Espada *et al.* 2021). Awareness of POD obliteration, regardless of risk for endometriosis, is relevant as it informs clinicians about the etiology of symptoms, guides medical and surgical treatments for pain and infertility, and provides vital information for surgical risk stratification (Brummer *et al.* 2011).

No radiology society yet recommends routine evaluation for POD obliteration in the assessment of female pelvic pathologies. Even in the context of endometriosis, most gynecologists are not seeing POD obliteration evaluated on TVS from their local radiology practices (Leonardi *et al.* 2020*c*) despite the recommendations by the International Deep Endometriosis Analysis (IDEA) group (Guerriero *et al.* 2016). There are some obstacles, which have likely limited the uptake of the sliding sign test. The organizational nature of ultrasound requires an operator, often a sonographer, and a physician, often a radiologist. Sonographers must learn how to perform the technique and simultaneously interpret what they are seeing to ensure correct performance and adequate acquisition of a video for final interpretation by the radiologist, who must also learn how to interpret video recordings of a dynamic test. Learning curve studies have been completed but these usually involve expert sonologists performing and

interpreting the sliding sign simultaneously (Tammaa *et al.* 2014, Leonardi *et al.* 2020*b*). Though there has been an increased uptake of advanced ultrasound by sonographers (Collins *et al.* 2019), limitations still remain.

We believe the routine integration of the sliding sign into the practice of gynecologic ultrasound is likely to occur. The Australasian Society of Ultrasound in Medicine (ASUM) have updated their guidelines on the performance of a gynecologic scan, including a recommendation to include the sliding sign (Australasian Society for Ultrasound in Medicine 2019). DL model could assist sonographers and radiologists when the sliding sign is more broadly adopted. Maximizing the potential of technology may even encourage more rapid implementation since barriers could be reduced. For example, with such a high PPV, radiologists may not need to review every video that is deemed normal as per the DL model (i.e. positive sliding sign). Emphasizing a high specificity means radiologists could focus on the cases that are classified as negative and if necessary, a human interpretation could overrule that of the DL model. We expect that the widespread introduction of the sliding sign into gynecologic ultrasound, fortified by this DL model, has the potential to significantly and positively impact patient care.

Specific to endometriosis, the development of this DL model may advance our ability to diagnose women non-invasively, yielding benefits such as a reduction in the delay to diagnosis (Hudelist *et al.* 2012), acknowledgment of symptoms, and optimizing access to care (As-Sanie *et al.* 2019).

Limitations and strengths

The prospective nature, relatively large sample size, use of high-quality gynecologic ultrasound equipment, and participation by two expert sonologists are the study's strengths. However, there are study limitations. The decision to standardize the collection of videos and interpretation by only two expert sonologists limited the total number of videos attainable to train the DL model. To account for this limitation, we used a relatively low capacity pre-trained model to avoid overfitting the training data. A larger training set would allow the use of a higher capacity model that could capture more of the variability present in the TVS videos and thus increase its diagnostic performance. Specifically, with a larger sample of videos depicting a negative sliding sign, there should be improvements in the specificity, ensuring that patients are not falsely reassured as having a normal, non-obliterated POD. Resizing the temporal and spatial resolutions of the TVS videos due to the high computational

requirements removed details of the videos, probably impacting the performance of our DL model.

As stated above, learning to perform the sliding sign and correctly record the video clip are necessary to ensure that the DL model can be adequately applied. In this study, two expert sonologists performed and recorded the videos. One potential limitation to the application of this new methodology is the crucial necessity to provide satisfactory training for examiners to adequately perform the sliding sign, otherwise, the real utility of the DL method would be seriously compromised. When applying AI to imaging interpretation, the data is the essential core, so it does not eliminate the need for obtaining it properly.

Another limitation in the study is that we did not have surgical data confirming the state of the POD. However, the diagnostic accuracy of sonologist-interpreted sliding sign is high (Nisenblat *et al.* 2016) and there is evidence of almost perfect interobserver agreement of expert sonologists interpreting offline videos of the sliding sign (Chiu *et al.* 2019). A study involving all patients that undergo surgery would be advantageous, but it will be limited by the high prevalence of pathology that necessitates the surgery in the first place. An ultrasound-only study allows for broader recruitment and representation.

Finally, the gynecologic-focused nature of the ultrasound practice where this study took place likely fosters a higher prevalence of endometriosis and higher quality sliding sign videos. Therefore, this study may be not exactly reproducible if the setting had a different prevalence of disease or less standardized approach to recording the sliding sign. As only one brand of the ultrasound machine was used (GE Healthcare), additional studies including equipment from different brands should be considered. The same concept should be applied to the ultrasound operators: a larger number and diversity of sonographers, radiologists, and sonologists should be considered.

Conclusions

In this study, we developed an accurate DL model that successfully classified TVS videos depicting the sliding sign as positive or negative. This DL model could help further disseminate the sliding sign test leading to an increased assessment of the POD and recognition of an obliterated POD, which has important diagnostic, surgical, and healthcare cost implications (Leonardi *et al.* 2019), particularly for those with endometriosis. This study may encourage further research on deep learning models in the non-invasive diagnosis of gynecologic pathology.

References

Altman D, Machin D, Bryant T & Gardner M 2013 *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*, 2nd ed. BMJ Books, Wiley London, UK

As-Sanie S, Black R, Giudice LC, Gray Valbrun T, Gupta J, Jones B, Laufer MR, Milspaw AT, Missmer SA, Norman A, *et al*. 2019 Assessing research gaps and unmet needs in endometriosis. *American Journal of Obstetrics and Gynecology* **221** 86–94. (https://doi.org/10.1016/j.ajog.2019.02.033)

Australasian Society for Ultrasound in Medicine 2019 Guidelines for the performance of a gynaecological scan. (available at: http://www.asum.com.au/newsite/Files/Documents/Policies/updated/D8_Policy.pdf)

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, Vet HCW de, *et al*. 2015 STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Clinical Chemistry* **61** 1446–1452. (https://doi.org/10.1373/clinchem.2015.246280)

Brummer TH, Jalkanen J, Fraser J, Heikkinen AM, Kauko M, Mäkinen J, Seppälä T, Sjöberg J, Tomás E & Härkki P 2011 FINHYST, a prospective study of 5279 hysterectomies: complications and their risk factors. *Human Reproduction* **26** 1741–1751. (https://doi.org/10.1093/humrep/der116)

Chiu LC, Leonardi M, Lu C, Mein B, Nadim B, Reid S, Ludlow J, Casikar I & Condous G 2019 Predicting pouch of douglas obliteration using ultrasound and laparoscopic video sets: an interobserver and diagnostic accuracy study. *Journal of Ultrasound in Medicine* **38** 3155–3161. (https://doi.org/10.1002/jum.15015)

Collins BG, Ankola A, Gola S & McGillen KL 2019 Transvaginal US of endometriosis: looking beyond the endometrioma with a dedicated protocol. *RadioGraphics* **39** 1549–1568. (https://doi.org/10.1148/rg.2019190045)

Cullen TS 1914 Adenomyoma of the rectovaginal septum. *JAMA* **LXII** 835. (https://doi.org/10.1001/jama.1914.02560360015006)

Drukker L, Sela HY, Reichman O, Rabinowitz R, Samueloff A & Shen O 2018 Sliding sign for intra-abdominal adhesion prediction before repeat cesarean delivery. *Obstetrics and Gynecology* **131** 529–533. (https://doi.org/10.1097/AOG.0000000000002480)

Espada M, Leonardi M, Aas-Eng K, Lu C, Reyftmann L, Tetstall E, Slusarczyk B, Ludlow J, Hudelist G, Reid S, *et al*. 2021 A multicenter international temporal and external validation study of the ultrasound-based endometriosis staging system. *Journal of Minimally Invasive Gynecology* **28** 57–62. (https://doi.org/10.1016/j.jmig.2020.04.009)

European Federation of Societies for Ultrasound in Medicine and Biology 2006 Minimum training recommendations for the practice of medical ultrasound. *Ultraschall der Medizin* **27** 79–105. (https://doi.org/10.1055/s-2006-933605)

Guerriero S, Condous G, Bosch T van den, Valentin L, Leone FPG, Schoubroeck D Van, Exacoustos C, Installé AJF, Martins WP, Abrao MS, *et al*. 2016 Systematic approach to sonographic evaluation of the pelvis in women with suspected endometriosis, including terms, definitions and measurements: a consensus opinion from the International Deep Endometriosis Analysis (IDEA) Group. *Ultrasound in Obstetrics and Gynecology* **48** 318–332. (https://doi.org/10.1002/uog.15955)

Hudelist G, Fritzer N, Thomas A, Niehues C, Oppelt P, Haas D, Tammaa A & Salzer H 2012 Diagnostic delay for endometriosis in Austria and Germany: causes and possible consequences. *Human Reproduction* **27** 3412–3416. (https://doi.org/10.1093/humrep/des316)

Hudelist G, Fritzer N, Staettner S, Tammaa A, Tinelli A, Sparic R & Keckstein J 2013 Uterine sliding sign: a simple sonographic predictor for presence of deep infiltrating endometriosis of the rectum. *Ultrasound in Obstetrics and Gynecology* **41** 692–695. (https://doi.org/10.1002/uog.12431)

Kingma DP & Ba J 2015 Adam: a method for stochastic optimization. (available at: http://arxiv.org/abs/1412.6980)

Leonardi M, Martin E, Reid S, Blanchette G & Condous G 2019 Deep endometriosis transvaginal ultrasound in the workup of patients with signs and symptoms of endometriosis: a cost analysis. *BJOG* **126** 1499–1506. (https://doi.org/10.1111/1471-0528.15917)

Leonardi M, Martins WP, Espada M, Georgousopoulou E & Condous G 2020*a* Prevalence of negative sliding sign representing pouch of douglas obliteration during pelvic transvaginal ultrasound for any indication. *Ultrasound in Obstetrics and Gynecology* **56** 928–933. (https://doi.org/10.1002/uog.22023)

Leonardi M, Ong J, Espada M, Stamatopoulos N, Georgousopoulou E, Hudelist G & Condous G 2020*b* One-size-fits-all approach does not work for gynecology trainees

learning endometriosis ultrasound skills. *Journal of Ultrasound in Medicine* **39** 2295–2303. (https://doi.org/10.1002/jum.15337)

Leonardi M, Robledo KP, Goldstein SR, Benacerraf BR & Condous G 2020c International survey finds majority of gynecologists are not aware of and do not utilize ultrasound techniques to diagnose and map endometriosis. *Ultrasound in Obstetrics and Gynecology* **56** 324–328. (https://doi.org/10.1002/uog.21996)

Melnyk A, Rindos NB, Khoudary SR El & Lee TTM 2020 Comparison of laparoscopic hysterectomy in patients with endometriosis with and without an obliterated cul-de-sac. *Journal of Minimally Invasive Gynecology* **27** 892–900. (https://doi.org/10.1016/j.jmig.2019.07.001)

Menakaya U, Infante F, Lu C, Phua C, Model A, Messyne F, Brainwood M, Reid S & Condous G 2016 Interpreting the real-time dynamic 'sliding sign' and predicting pouch of douglas obliteration: an interobserver, intraobserver, diagnostic-accuracy and learning-curve study. *Ultrasound in Obstetrics and Gynecology* **48** 113–120. (https://doi.org/10.1002/uog.15661)

Mercaldo ND, Lau KF & Zhou XH 2007 Confidence intervals for predictive values with an emphasis to case–control studies. *Statistics in Medicine* **26** 2170–2183. (https://doi.org/10.1002/sim.2677)

Nisenblat V, Bossuyt PMM, Farquhar C, Johnson N & Hull ML 2016 Imaging modalities for the non-invasive diagnosis of endometriosis. *Cochrane Database of Systematic Reviews* **2** CD009591. (https://doi.org/10.1002/14651858.CD009591.pub2)

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, *et al*. 2019 PyTorch: an imperative style, high-performance deep learning library. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. (available at: http://arxiv.org/abs/1912.01703)

Purohit R, Sharma JG, Meher D, Rakh SR & Malik S 2018 Completion of vaginal hysterectomy by electro surgery using anteroposterior approach in benign cases faced with obliterated posterior cul-de-sac. *International Journal of Women's Health* **10** 529–536. (https://doi.org/10.2147/IJWH.S171575)

Reid S, Lu C, Casikar I, Reid G, Abbott J, Cario G, Chou D, Kowalski D, Cooper M & Condous G 2013 Prediction of pouch of douglas obliteration in women with suspected endometriosis using a new real-time dynamic transvaginal ultrasound technique: the sliding sign. *Ultrasound in Obstetrics and Gynecology* **41** 685–691. (https://doi.org/10.1002/uog.12305)

Tammaa A, Fritzer N, Strunk G, Krell A, Salzer H & Hudelist G 2014 Learning curve for the detection of pouch of douglas obliteration and deep infiltrating endometriosis of the rectum. *Human Reproduction* **29** 1199–1204. (https://doi.org/10.1093/humrep/deu078)

Tompsett J, Leonardi M, Gerges B, Lu C, Reid S, Espada M & Condous G 2019 Ultrasound-based endometriosis staging system: validation study to predict complexity of laparoscopic surgery. *Journal of Minimally Invasive Gynecology* **26** 477–483. (https://doi.org/10.1016/j.jmig.2018.05.022)

Tran D, Wang H, Torresani L, Ray J, LeCun Y & Paluri M 2018 A closer look at spatiotemporal convolutions for action recognition. In *Proceedings – 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459. (https://doi.org/10.1109/CVPR.2018.00675)