Systematic review of research design and methodology quality of studies applying deep learning in ultrasound foetal head circumference measurements.

Author: Luis Aparicio, MSc in Medical Ultrasound.
University of the West of England.
Yeovil District Hospital NHS Foundation Trust

Abstract

Objectives This systematic review was conducted to evaluate the research design and reporting in articles assessing the accuracy of deep learning algorithms in the measurement of the foetal head during an ultrasound examination.

Methods Electronic literature searches were carried out in Medline, CINAHL, EMBASE, ArXiv, IEEE and Scopus. As this review was meant to capture all relevant literature, all studies reporting on the accuracy of 2D ultrasound measurements of the foetal head using deep learning algorithms, with a clear description of their architecture, were included.

Results All studies included in this systematic review show some degree of hidden bias, mainly due to dataset selection and/or the omission or exclusion of important steps in the model training.

Conclusions Although the implementation of newly developed guidelines has already shown improved reporting standards of new studies, more care is needed in model training and construction of databases.

Key Points

- Deep learning studies segmenting and measuring the foetal head show high levels of accuracy but have critical flaws in their design and show high level of hidden bias.
- Study designs have improved since the release of new artificial intelligence specific guidelines.
- Most hidden bias seem to stem from how the algorithm training is documented and database creation.
- Further care is needed when creating databases and future studies should directly compare machine vs. human performance.

Keywords
Ultrasound, Deep Learning, Foetal Head Circumference, Automated, Accuracy, Study Design, Risk of Bias.

Abbreviations and acronyms:

AC – Abdomen Circumference
AI – Artificial Intelligence
BMI – Body Mass Index
CLAIM – Checklist for AI in Medical Imaging
CNN – Convolutional neural network
CONSORT-AI - Consolidated Standards of Reporting Trials-Artificial Intelligence
DL – Deep Learning
EFW – Estimated Foetal Weight
EQUATOR – Enhancing the Quality and Transparency of Health Research
FL – Femur Length
HC – Head Circumference
IMRaD – Introduction, Methods, Results and Discussion
QUADAS - The Quality Assessment of Diagnostic Accuracy Studies
PRISMA – Preferred Reporting Items for Diagnostic Accuracy Studies
SPIRIT-AI – Standard Protocol Items: Recommendations for Interventional Trials Artificial Intelligence extension
STARD – Standards for Reporting of Diagnostic Accuracy Studies
SPIRIT-AI – Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence

Introduction

In the second and third trimester of pregnancy, it is common practice to use ultrasound to measure the foetal head (head circumference - HC), abdomen (abdominal circumference - AC) and the femur (femur length - FL). These measurements can be combined in established formulae in order to obtain an estimated foetal weight (EFW) and help predict foetal growth disorders [1]. Many factors can impact on the quality of these measurements. These may include foetal position, increased maternal body mass index (BMI) and placental position [2, 3]. Foetal growth measurements can also vary when performed by different sonographers (inter-observer variability) and even when performed by the same sonographer (repeatability of measurements - intraobserver variability). Caliper placement was shown to be the main component of overall variability in foetal growth measurements (inter- and intra- variance), and it is mainly influenced by the skill and level of experience of the sonographer [4]. However, factors such as pressures from busy clinical practice and sonographer fatigue should also be considered [2, 3, 5, 6]. To fix this problem, artificial intelligence (AI) systems have been developed and researched with the goal of automating these measurements and therefore reduce variability [7–10]. Early developments in this field were only capable of basic pattern recognition outcomes; however, later improvements, where complex and more specialised machine learning algorithms were developed, allowed researchers to achieve very promising results in image segmentation [5, 7, 8, 11]. Although these machine learning methods were promising, they process information based on rigid "hand crafted" parameters, composed by engineers and therefore may lack the ability to generalise and extrapolate and may be subjective [12–14]. For example, a machine learning algorithm may be very accurate in identifying an image of an apple, based on rigid parameters such as shape and colour, but may fail to correctly label a rotten apple as an apple due to the possible change in shape or colour. A more robust approach to this problem is to develop algorithms capable of inferring new knowledge by analysing large quantities of labelled data and be able to decide on parameters and exceptions, on its own, based on statistics and probability [9, 10]. This method is a subset of machine learning, where multiple layers of algorithms (neural networks) are utilised, called Deep Learning (DL). The current most successful neural network for computer vision is called convolutional neural networks (CNN) [14, 15].

The benefit of implementing an automated system to perform time consuming and repetitive tasks is clear and many studies using different CNN architectures have reported very promising results, specifically in the segmentation and measurement of the foetal head. However, due to their specific nature, these studies introduce challenges in the way their accuracy and reliability is reported. Traditional guidelines may not be adequate to fully uncover important information that may be missed and increase risk of hidden bias [16]. A recent systematic review [17] studying the diagnostic accuracy of deep learning compared to health-care professionals, highlighted that more than 99% of the 20,500 articles examined showed inadequate design and/or reporting. Although this paper focused on pathology classification rather than image segmentation, it is probable that similar results would have been observed.

This finding was most likely due to the lack of specifically customised guidelines in the development and reporting of articles assessing the accuracy of AI algorithms in healthcare. Efforts were made in order to mitigate these limitations with AI specific extensions being recently published, such as the SPIRIT-AI[18] (Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence) and CONSORT-AI[19] (Consolidated Standards of Reporting Trials-Artificial Intelligence). Other new guides and guidelines were also developed with further specific attention to AI studies in medical imaging [20, 21]. The adoption of an automated method capable of performing foetal growth measurements in clinical practice requires careful and rigorous confirmation of its clinical reliability and usefulness [22]. Therefore, the aim of this systematic review is to identify studies reporting the accuracy of deep learning algorithms in the measurement of the foetal head and evaluate their design and methodology quality using recent AI specific guidelines. Only this way can evidence based practice be achieved.

Materials and Methods

The protocol used to conduct this review was the Preferred Reporting Items of Systematic Reviews and Meta-Analysis (PRISMA) [23].

Eligibility criteria in literature search

All studies reporting on the accuracy of DL algorithms in the measurement of the foetal head and appropriately showed its internal architecture were eligible. DL algorithms that used 3D volumes to segment the foetal head were excluded from this review.

Information sources

MEDLINE, Embase, ArXiv and CINAHL were the main databases utilised. Reference lists of included articles were also manually scrutinised in order to obtain missed relevant articles. Grey literature included conference abstract/paper proceedings by searching IEEE. A reverse tracing strategy was also utilised by using titles of included articles in the Scopus database in order to identify additional papers that referenced the included article.

Keyword terms in four different categories (PICO) as well as synonyms and Medical subject Headings (MeSH) were utilised and combined with Boolean operators to maximise potential article recall. The detailed search strategy can be found in the supplementary material.

Selection process

The primary selection was centred around articles that clearly described the deep learning algorithm utilised to obtain a 2D foetal head circumference measurement, so that the reviewer could, at least, understand how the image segmentation is processed. Many articles report on machine brand specific methods, mainly by referring to them by their commercial name, without specifying internal architecture. These studies were excluded. The selection and collection of data was performed by only one person, the author, and the process is illustrated by the PRISMA flow chart (Figure 1).

Data extraction

Data was extracted in view of the study design checklist (Table 1). This checklist was constructed by identifying appropriate signalling questions based on current checklists and guidelines in the evaluation of deep learning methods in medical imaging [20–22]. These questions were mainly influenced by the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) [20] as this was specifically designed to help authors and reviewers of AI manuscripts in medical imaging and was modelled after STARD (Standards for Reporting of Diagnostic Accuracy Studies) [25]. Due to intricacies of the different DL methods and their applications to different tasks in medical imaging, it is difficult to adopt a checklist or guideline that can eliminate all sources of possible bias. For example, in the obstetric field, the exclusion of patients with high BMI, or pregnancies in the third trimester could overinflate the accuracy of growth measurements of a DL algorithm. Therefore, extra attention was taken in the selection of appropriate questions to specifically evaluate the datasets included in the different studies.

Table 1: Study design checklist

	Item	Criterion	Values	Compliant values
Title/Abstract	1	Identification as a study of AI methodology, specifying the category of technology used (e.g., deep learning)	0. Not specified 1. Specified	1
	2	Structured summary of study design, methods, results, and conclusions	0. Not specified 1. Specified	1
Introduction	3	Scientific and clinical background, including the intended use and clinical role of the AI approach	0. Not specified 1. Specified	1
	4	Study objectives and hypotheses	0. Not specified 1. Specified	1
Methods	5	Prospective or retrospective study	0. Not specified R. Retrospective P. Prospective	R, P
	6	Study goal, such as model creation, exploratory study, feasibility study, non-inferiority trial	0. Not specified 1. Specified	1
	7a	Data sources	0. Not specified 1. Local data 2. Public data	1, 2
	7b	Different stages of pregnancy included	0. Not specified 1. Not Appropriate 2. Appropriate	2
	7c	Appropriate of images with artifacts and/or difficult to obtain due to foetal position.	0. Not specified 1. Not Appropriate 2. Appropriate	2
	7d	Appropriate of images in the later stages of pregnancy.	0. Not specified 1. Not Appropriate 2. Appropriate	2
	7e	BMI of patients	0. Not specified 1. Specified	1
	7f	Multiple ultrasound machine vendors	0. Not specified 1. Single vendor 2. Multiple vendors	2
	8	Eligibility criteria: how, where, and when potentially eligible participants or studies were identified (e.g., symptoms, results from previous tests, inclusion in registry, patient-care setting, location, dates)	0. Not specified 1. Limited 2. Detailed	1,2
	9	Data pre-processing steps	0. Not specified 1. Specified	1
	10	De-identification methods	0. Not specified 1. Specified	1
	11	Definition of ground truth reference standard, in sufficient detail to allow replication	0. Not specified 1. Specified	1
	12	Source of ground-truth annotations; qualifications and preparation of annotators	0. Not specified 1. Specified	1
	13a	Intended sample size	0. Not specified 1. Specified	1
	13b	Rational for sample size	0. Not specified 1. Specified	1
	13c	How data were assigned to partitions	0. Not specified 1. Specified	1
	13d	Specify proportions	0. Not specified 1. Specified	1
	14	Level at which partitions are disjoint (e.g., image, study, patient, institution)	0. Not specified 1. Specified	1
	15	Detailed description of model, including inputs, outputs, all intermediate layers and connections	0. Not specified 1. Limited 2. Detailed	1, 2
	16	Software libraries, frameworks, and packages	0. Not specified 1. Specified	1
	17	Initialization of model parameters (e.g., randomization, transfer learning)	0. Not specified 1. Limited 2. Detailed	1, 2
	18a	Details of training approach, including data augmentation, hyperparameters, number of models trained	0. Not specified 1. Limited 2. Detailed	1, 2
	18b	Distinction between training and validation	0. Not specified 1. Specified	1
	19	Metrics of model performance	0. Not specified 1. Specified	1
	20	Statistical measures of significance and uncertainty (e.g., confidence intervals)	0. Not specified 1. Specified	1
	21	Robustness or sensitivity analysis	0. Not specified 1. Specified	1
	22	Validation or testing on external data	0. Not specified 1. Temporal testing 2. Geographical testing	1, 2
Results	23	Flow of participants or cases, using a diagram to indicate inclusion and exclusion	0. Not specified 1. Specified	1
	24	Demographic and clinical characteristics of cases in each partition	0. Not specified 1. Specified	1
	25	Performance metrics for optimal model(s) on all data partitions	0. Not specified 1. Specified	1
	26	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	0. Not specified 1. Specified	1
Discussion	27	Study limitations, including potential bias, statistical uncertainty, and generalizability	0. Not specified 1. Specified	1
	28	Implications for practice, including the intended use and/or clinical role	0. Not specified 1. Specified	1

Data analysis

All items in the study design checklist (Table 1) have one or two compliant values. Compliance ratios were calculated by dividing the number of items satisfied by the total number of applicable items. **Table 2** breaks down **Table 1** into the IMRaD (Introduction, Methods, Results and Discussion) structure of each paper and highlights the compliance in each section. For example: A compliance of 12/27 means only 12 items were satisfied out of 27 total items. An overall total compliance percentage was also given for each study.

The Quality Assessment of Diagnostic Accuracy Studies – 2 (QUADAS-2) [26] tool was also utilised for easier presentation of possible methodological quality limitations. This is achieved by assessing four different domains: patient selection, index test, reference standard, flow and timing. Each domain is attributed a high, unclear or low risk of bias assessment. Three further trigger questions were also included in the three first domains for the assessment of concerns regarding applicability. A traffic light plot was used to illustrate the risk of bias in the design of the different studies (Figure 2). The complete risk of bias assessment is available in the supplementary material.

Results

Study selection

A total of 365 studies were identified through database search. After removal of duplicates the total of articles was reduced to 255. The title of these 255 articles were assessed and 181 articles were excluded (mainly due to no use of artificial intelligence at all or only use of machine learning algorithms). The abstracts of the remaining papers were further scrutinised against the exclusion criteria and 37 articles were excluded. After assessing the full texts of the remaining papers, 29 articles were again excluded due to only reporting image segmentation with no measurement (n = 6), anatomy identification only (n = 8), no clear method of image segmentation (n= 12), pathology specific dataset (patients with polyhydramnios) (n = 1), and 3D image dataset only (n = 2). Five articles were further identified and included in the final review by reverse searching included studies using SCOPUS. The final number of studies included in this review was 13.

Table 2: Study design compliance checklist

	Study design compliance checklist													
Author	Wu et al.[27]	Sinclair et al.[36]	Rong et al.[28]	Oghli et al.[31]	Sobhaninia et al.[32]	Al-Bander et al.[37]	Li et al.[29]	Zhang et al.[30]	Qiao et al.[38]	Oghli et al.[33]	Fiorentino et al.[34]	Moccia et al.[35]	Zeng et al.[10]	Average
Title/Abstract	2/2	2/2	2/2	1/2	2/2	2/2	2/2	2/2	2/2	2/2	2/2	2/2	2/2	96%
Introduction	2/2	2/2	2/2	1/2	2/2	2/2	2/2	2/2	2/2	2/2	2/2	2/2	2/2	96%
Methods	12/27	17/27	19/27	16/27	21/27	20/27	20/27	13/27	21/27	20/27	22/27	22/27	22/27	69%
Results	2/4	2/4	2/4	2/4	2/4	2/4	2/4	2/4	2/4	2/4	2/4	2/4	2/4	50%
Discussion	1/2	1/2	1/2	1/2	1/2	1/2	2/2	1/2	2/2	1/2	2/2	2/2	2/2	69%
Total	19/37 - 51%	24/37 - 65%	26/37 - 70%	21/37 57%	28/37 - 76%	27/37 - 73%	28/37 - 76%	20/37 - 54%	29/37 - 78%	27/37 - 73%	30/37 - 81%	30/37 - 81%	30/37 - 81%	70%

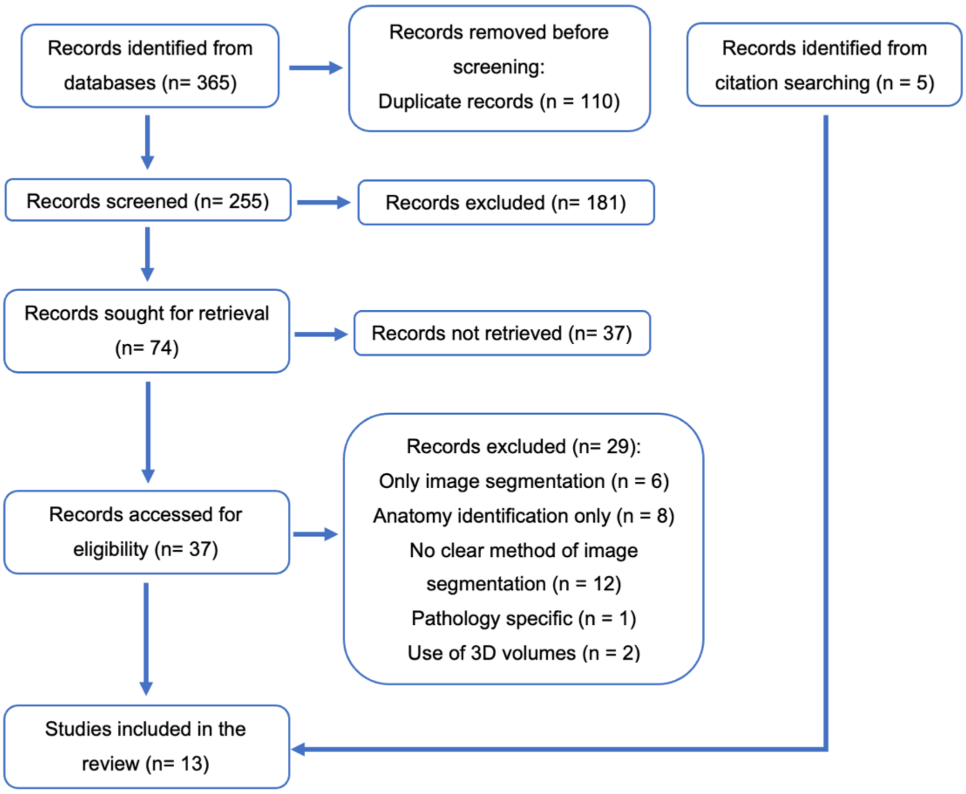

Records identified from databases (n= 365)
Records removed before screening:
Duplicate records (n = 110)
Records identified from citation searching (n = 5)
Records screened (n= 255)
Records excluded (n= 181)
Records sought for retrieval (n= 74)
Records not retrieved (n= 37)
Records excluded (n= 29):
Only image segmentation (n = 6)
Anatomy identification only (n = 8)
No clear method of image segmentation (n = 12)
Pathology specific (n = 1)
Use of 3D volumes (n = 2)
Records accessed for eligibility (n= 37)
Studies included in the review (n= 13)

Figure 1: PRISMA flow diagram

Study characteristics

Five out of the thirteen studies included are from China [10, 27–30], three from Iran [31–33], two from Italy [34, 35] and one from England [36], Iraq [37] and Canada [38]. When scrutinised against the study design checklist, the included studies showed an overall compliance of 70%. The highest compliance rate was 81% [10, 34, 35] and the lowest 51% [27]. (See table 2)

With one exception [31], all studies identified their research as studies of AI methodology and showed structured summary of design, methods, results and conclusions.

Good overall compliance was also achieved in the introduction section of the studies with only one study [31] not including a clinical background component.

In the methods section, 92% of the studies stated their data sources and 85% specified the use of the three different trimesters of pregnancy. Eleven out of thirteen studies used publicly available datasets to train and test their algorithms. Zhang et al. [30] utilised the dataset introduced by Ryou et al. [39]. The remaining studies utilised the dataset published by Heuvel et al. [40]. Only one study [36] used own data, gathered retrospectively.

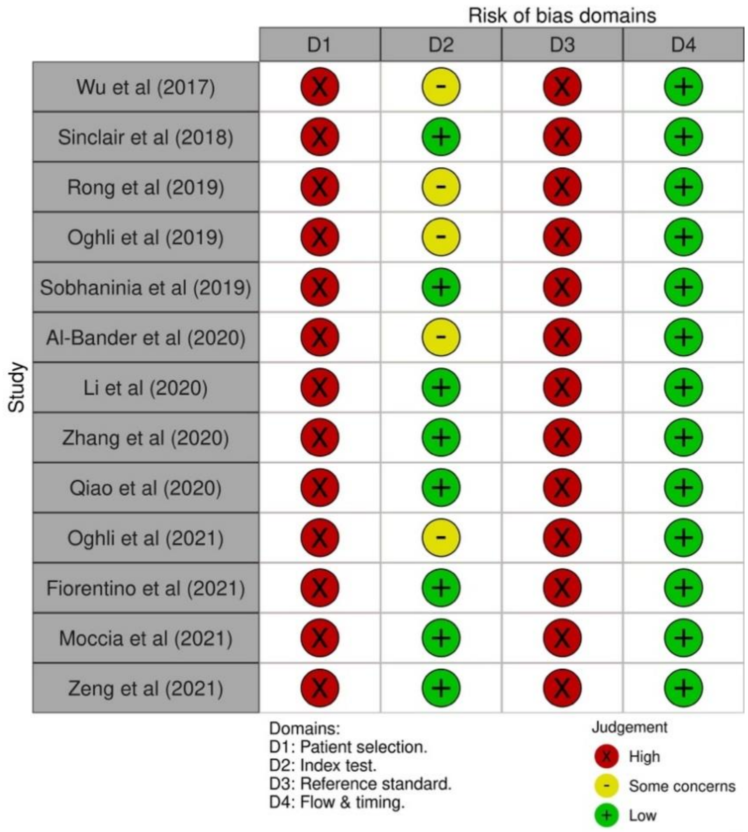
All studies showed a discrepancy in the number of images in the different stages of pregnancy in favour of the second trimester. No studies specified patient characteristics and the only exclusion criteria mentioned was the active exclusion of patients with growth abnormalities. Two studies [30, 36] did not specify any exclusion criteria.

All studies showed detailed description of their DL models, including inputs, outputs, intermediate layers and connections. Furthermore, only 69% specified software libraries, frameworks and packages used. A distinction between training and validation was only observed in 46% of the studies. All studies specified their sample size; however, none stipulated a rational for this number or how the specific number of pictures were assigned to each partition.

Ten out of thirteen studies reported use of image augmentation to generate extra images for training. However, only two of these studies specified the number of extra images generated [10, 32].

All studies presented their results using appropriate performance metrics and estimates of diagnostic accuracy. None of the studies included in this review utilised diagrams to illustrate their inclusion or exclusion criteria or patient characteristics.

In the discussion section, only 39% of the studies included study limitations and potential bias. The complete study design checklist table is available in the supplementary material.



Risk of bias domains

Study	D1	D2	D3	D4
Wu et al (2017)	X	-	X	+
Sinclair et al (2018)	X	+	X	+
Rong et al (2019)	X	-	X	+
Oghli et al (2019)	X	-	X	+
Sobhaninia et al (2019)	X	+	X	+
Al-Bander et al (2020)	X	-	X	+
Li et al (2020)	X	+	X	+
Zhang et al (2020)	X	+	X	+
Qiao et al (2020)	X	+	X	+
Oghli et al (2021)	X	-	X	+
Fiorentino et al (2021)	X	+	X	+
Moccia et al (2021)	X	+	X	+
Zeng et al (2021)	X	+	X	+

Domains:
D1: Patient selection.
D2: Index test.
D3: Reference standard.
D4: Flow & timing.

Judgement
X High
- Some concerns
+ Low

Figure 2: Risk of bias checklist

Discussion

The accuracy of DL algorithms in the measurement of the foetal head is very promising and show at least comparable performance to the human experts [36]. Moreover, another recent study [5] detailed how the implementation of an automated algorithm in the measurement of the foetal head, significantly improved workflow by reducing the measurement time by 61% and cutting measuring steps in half. All these results are very optimistic, however, and further care is needed in order to ascertain their validity in real medical practice.

Some of the studies included in this review were developed prior to the elaboration of AI specific guidelines and therefore show some crucial flaws in their design; however, this review shows a significant improvement in the reporting standards after the release of AI adapted checklists. This finding was also reported by O’Shea et al. [41]. There is, however, room for improvement. The risk of bias assessment highlights high levels of possible bias in the index test, reference test and patient selection. The main reasons for this assessment seem to stem from two different components of the study design: Model training and dataset selection.

Model training

Although most studies do a good job presenting their models and internal architecture, one important area to improve is when presenting data referring to the model training. To effectively train and test a DL model, the dataset must be divided in three different partitions: training, validation and testing. The training subset is used for model learning, the validation dataset is used to fine tune the algorithm and for model selection and the testing subset is used to assess performance of a finalised model [41]. Approximately half of the studies in this review only reported two data partitions. Therefore, their results may well be only a representation of the performance of the training or validation stages. If this is the case, these results only show the internal validity of the algorithm and misrepresent generalisability metrics optimistically [41].

Dataset selection

The construction of a dataset for a DL algorithm for medical diagnosis/prediction poses unique challenges. Although most of the studies utilised the HC18 Grand Challenge dataset [40] which is considered as benchmark in the field, there are several limitations:

- Patient characteristics - There is very limited information concerning the patients included in the dataset. It is mentioned that no growth restricted patients were included which may cause spectrum bias, a phenomenon that occurs when the sample population is not representative of the target population. Unlike in prospective clinical trials, where patients are selected consecutively following a previously established eligibility criteria, the data used to train a DL algorithm must be collected from multiple heterogenous sources and be representative of the target patient population [22]. No information was also given on the average BMI of the patients included. This is a particularly important point especially due to the effect that BMI has on EFW [42].
- Inclusion of all trimesters of pregnancy – This dataset includes all stages of pregnancy however there is a discrepancy in the number of images in the three

different trimesters. While 70% of the images in the dataset refer to the second trimester of pregnancy, only 14% are from the third trimester and 16% from the first trimester. Many studies [43–45] have shown how the accuracy of the EFW decreases as the gestational age increases. Furthermore, Zeng et al. [10] reported on the difficulty in the segmentation of the foetal head in the first trimester of pregnancy due to lack of skull contour and increased noise, resulting in less accurate measurements. Although it is stated that difficult to obtain images and images with artefacts and with no clear borders were included in the dataset, to represent real-world practice, this obvious discrepancy in favour of images in the second trimester may skew the data into more favourable measurements and overestimate the accuracy of the tested models.

- Rational for number of images in the dataset – Although the sample size is described, there is no information on how it was determined. Mongan et al. [20] advises the use of traditional power calculation methods to estimate the sample size in order to ascertain its generalisability in a larger population.
- External validation/testing – The HC18 dataset has specified subsets for training and testing. This is usually classed as a split-sample model validation. However, the authors maintain that all images obtained during a single examination were all kept together, either in the training or the testing dataset. This avoids overfitting of the model, a situation where a model over-customises itself to the training dataset, achieving very good results during the training phase, but underperforms in the testing phase due to lack of generalisability to new data [22]. Although this kind of validation/testing is preferable to internal validation, where the testing is performed with images from the training phase, a more robust and generalisable model must be tested on completely external datasets. These should be acquired by external researchers, in a different setting, with different machine makes and different patient populations [22]. Only this way a model performance can be extrapolated to real-world clinical practice.

Conclusion

The potential of DL algorithms in foetal biometry is evident; however, all studies in this review showed some flaws in their design and lacked standardisation. This was most likely due to the lack of specialised guidelines. Since the recent release of updated guidelines in the reporting of AI algorithms in clinical practice, there has been a significant shift in the reporting of studies, showing better designs and therefore decreasing the risk of bias. More work is needed in the development of datasets that can remove any hidden bias and represent real-world practice. At minimum, the testing should be performed with high rigour, with completely external data in order to verify generalisability. Further studies should endeavour to test the accuracy of both algorithms and human experts using the same datasets. Only then can the results be extrapolated to clinical practice.

As it stands, there is no clear evidence to suggest that current deep learning algorithms can perform foetal head measurements autonomously, in real clinical practice.

Limitations

The process of data selection, extraction and quality assessment was performed by only one person (the author). This may constitute a limitation due to possible reviewer selection bias. Although the somewhat rigid eligibility criteria and transparent search strategy should mitigate some bias, there is the possibility that if more researchers were involved, other search strategies could have been implemented and more relevant studies might have been retrieved. Also, due to time constraints, the authors of the different studies included in this review were not contacted in order to obtain extra details about their studies that could reduce the bias observed. There is the possibility that some of these studies were conducted in a more precise manner and important details were inadvertently not conveyed in the final published paper.

Declarations

Conflict of interest - The author of this manuscript declares no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry - No complex statistical methods were necessary for this paper.

Informed consent - Written informed consent was not required for this study because this was a systematic review using published studies in the literature but not analysing specific human subjects.

Ethical approval - This research is exempt from ethics approval given that this is a systematic review, which uses published data that, subsequently, have been ethically approved.

References

1. Hadlock FP, Harrist RB, Sharman RS, Deter RL, Park SK (1985) Estimation of fetal weight with the use of head, body, and femur measurements—A prospective study. American Journal of Obstetrics & Gynecology 151:333–337. https://doi.org/10.5555/uri:pii:0002937885902984
2. Krispin E, Dreyfuss E, Fischer O, Wiznitzer A, Hadar E, Bardin R (2020) Significant deviations in sonographic fetal weight estimation: causes and implications. Arch Gynecol Obstet 302:1339–1344
3. Hendin N, Levin G, Tsur A, Ilan H, Rottenstreich A, Meyer R Factors Associated with More Than 500 Grams Inaccuracy in Sonographic Fetal Weight Estimation. The Israel Medical Association journal : IMAJ JID - 100930740
4. Sarris I, Ioannou C, Chamberlain P, Ohuma E, Roseman F, Hoch L, Altman DG, Papageorghiou AT, (INTERGROWTH-21st) IF and NGC for the 21st C (2012) Intra- and interobserver variability in fetal ultrasound measurements. Ultrasound in obstetrics & gynecology 39:266–273
5. Espinoza J, Good S, Russell E, Lee W (2013) Does the Use of Automated Fetal Biometry Improve Clinical Work Flow Efficiency? Journal of Ultrasound in Medicine 32:847–850. https://doi.org/https://doi.org/10.7863/jum.2013.32.5.847
6. Rottenstreich A, Yanai N, Yagel S, Porat S The Accuracy of Sonographic Assessment for Fetal Weight: Technicians versus Ultrasound-Certified Physicians. The Israel Medical Association journal : IMAJ JID - 100930740
7. Chalana V, Winter TC 3rd, Cyr DR, Haynor DR, Kim Y (1996) Automatic fetal head measurements from sonographic images. Acad Radiol 3:628–635
8. Carneiro G, Georgescu B, Good S, Comaniciu D (2008) Detection and measurement of fetal anatomies from ultrasound images using a constrained probabilistic boosting

tree. IEEE Trans Med Imaging 27:1342–1355. https://doi.org/https://dx.doi.org/10.1109/TMI.2008.928917

9. Kim HP, Lee SM, Kwon J-Y, Park Y, Kim KC, Seo JK (2019) Automatic evaluation of fetal head biometry from ultrasound images using machine learning. Physiol Meas 40:065009. https://doi.org/https://dx.doi.org/10.1088/1361-6579/ab21ac

10. Zeng Y, Tsui P-H, Wu W, Zhou Z, Wu S (2021) Fetal Ultrasound Image Segmentation for Automatic Head Circumference Biometry Using Deeply Supervised Attention-Gated V-Net. Journal of Digital Imaging 34:134–148. https://doi.org/10.1007/s10278-020-00410-5

11. Benjamin AEAFAU-A, Amoah B FAU - Crimi A, Crimi A (2018) Segmentation of ultrasound images of fetal anatomic structures using random forest for low-cost settings. Annual International Conference of the IEEE Engineering in Medicine and Biology Society.IEEE Engineering in Medicine and Biology Society.Annual International Conference JID - 101763872

12. Erickson BJ, Korfiatis P, Akkus Z, Kline TL (2017) Machine Learning for Medical Imaging. RadioGraphics 37:505–515. https://doi.org/10.1148/rg.2017160130

13. Brattain LJ, Telfer BA, Dhyani M, Grajo JR, Samir AE (2018) Machine learning for medical ultrasound: status, methods, and future opportunities. Abdominal radiology (New York) JID - 101674571

14. Drukker L, Noble JA, Papageorghiou AT (2020) Introduction to artificial intelligence in ultrasound imaging in obstetrics and gynecology. Ultrasound Obstet Gynecol 56:498–505. https://doi.org/10.1002/uog.22122

15. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88. https://doi.org/https://doi.org/10.1016/j.media.2017.07.005

16. Shelmerdine SC, Arthurs OJ, Denniston A, Sebire NJ (2021) Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare. BMJ Health Care Inform 28:e100385. https://doi.org/10.1136/bmjhci-2021-100385

17. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, Mahendiran T, Moraes G, Shamdas M, Kern C (2019) A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The lancet digital health 1:e271–e297

18. Rivera SC, Liu X, Chan A-W, Denniston AK, Calvert MJ (2020) Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. bmj 370:

19. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Chan A-W, Darzi A, Holmes C, Yau C, Ashrafian H, Deeks JJ, Ferrante di Ruffano L, Faes L, Keane PA, Vollmer SJ, Lee AY, Jonas A, Esteva A, Beam AL, Chan A-W, Panico MB, Lee CS, Haug C, Kelly CJ, Yau C, Mulrow C, Espinoza C, Fletcher J, Paltoo D, Manna E, Price G, Collins GS, Harvey H, Matcham J, Monteiro J, ElZarrad MK, Ferrante di Ruffano L, Oakden-Rayner L, McCradden M, Keane PA, Savage R, Golub R, Sarkar R, Rowley S, Group TS-A and C-AW, Group S-A and C-AS, Group S-A and C-AC (2020) Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nature Medicine 26:1364–1374. https://doi.org/10.1038/s41591-020-1034-x

20. Mongan J, Moy L, Kahn Jr CE (2020) Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers

21. Bluemke DA, Moy L, Bredella MA, Ertl-Wagner BB, Fowler KJ, Goh VJ, Halpern EF, Hess CP, Schiebler ML, Weiss CR (2020) Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers—from the radiology editorial board

22. Park SH, Han K (2018) Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. Radiology 286:800–809. https://doi.org/10.1148/radiol.2017171920

23. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372:n71. https://doi.org/10.1136/bmj.n71

24. Methley AM, Campbell S, Chew-Graham C, McNally R, Cheraghi-Sohi S (2014) PICO, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. BMC Health Serv Res 14:579. https://doi.org/10.1186/s12913-014-0579-0

25. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet HCW (2015) STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Clin Chem 61:1446–1452

26. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MMG, Sterne JAC, Bossuyt PMM, Group* Q-2 (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 155:529–536

27. Wu L, Xin Y, Li S, Wang T, Heng P-A, Ni D (2017) Cascaded fully convolutional networks for automatic prenatal ultrasound image segmentation. In: 2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017). IEEE, pp 663–666

28. Rong Y, Xiang D, Zhu W, Shi F, Gao E, Fan Z, Chen X (2019) Deriving external forces via convolutional neural networks for biomedical image segmentation. Biomedical optics express JID - 101540630 PMC - PMC6701547 COIS- The authors declare that there are no conflicts of interest related to this article.EDAT- 2019/08/28 06:00 MHDA- 2019/08/28 06:01 CRDT- 2019/08/28 06:00 PHST- 2019/05/15 00:00 rece(TRUNCATED

29. Li P, Zhao H, Liu P, Cao F (2020) Automated measurement network for accurate segmentation and parameter modification in fetal head ultrasound images. Med Biol Eng Comput 58:2879–2892. https://doi.org/https://dx.doi.org/10.1007/s11517-020-02242-5

30. Zhang L, Zhang J, Li Z, Song Y (2020) A multiple-channel and atrous convolution network for ultrasound image segmentation. Med Phys 47:6270–6285. https://doi.org/https://dx.doi.org/10.1002/mp.14512

31. Oghli MG, Shabanzadeh A, Moradi S, Gerami R (2019) Automatic fetal biometry evaluation in ultrasound images using a deep learningbased approach. Iranian Journal of Radiology 16:S11–S13. https://doi.org/http://dx.doi.org/10.5812/IRANJRADIOL.99141

32. Sobhaninia Z, Rafiei S, Emami A, Karimi N, Najarian K, Samavi S, Soroushmehr SMR (2019) Fetal Ultrasound Image Segmentation for Measuring Biometric Parameters Using Multi-Task Deep Learning. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp 6545–6548

33. Ghelich Oghli M, Shabanzadeh A, Moradi S, Sirjani N, Gerami R, Ghaderi P, Sanei Taheri M, Shiri I, Arabi H, Zaidi H (2021) Automatic fetal biometry prediction using a novel deep convolutional network architecture. Phys Med 88:127–137. https://doi.org/https://dx.doi.org/10.1016/j.ejmp.2021.06.020

34. Fiorentino MC, Moccia S, Capparuccini M, Giamberini S, Frontoni E (2021) A regression framework to head-circumference delineation from US fetal images. Comput Methods Programs Biomed 198:105771. https://doi.org/https://dx.doi.org/10.1016/j.cmpb.2020.105771

35. Moccia S, Fiorentino MC, Frontoni E (2021) Mask-R2CNN: a distance-field regression version of Mask-RCNN for fetal-head delineation in ultrasound images. International Journal of Computer Assisted Radiology and Surgery 16:1711–1718. https://doi.org/10.1007/s11548-021-02430-0

36. Sinclair M, Baumgartner CF, Matthew J, Bai W, Martinez JC, Li Y, Smith S, Knight CL, Kainz B, Hajnal J, King AP, Rueckert D (2018) Human-level Performance On Automatic Head Biometrics In Fetal Ultrasound Using Fully Convolutional Neural Networks. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp 714–717

37. Al-Bander B, Alzahrani T, Alzahrani S, Williams BM, Zheng Y (2020) Improving Fetal Head Contour Detection by Object Localisation with Deep Learning. In: Zheng Y, Williams BM, Chen K (eds) Medical Image Understanding and. Springer International Publishing, Cham, pp 142–150

38. Qiao D, Zulkernine F (2020) Dilated Squeeze-and-Excitation U-Net for Fetal Ultrasound Image Segmentation. In: 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). pp 1–7

39. Ryou H, Yaqub M, Cavallaro A, Papageorghiou AT, Alison Noble J (2019) Automated 3D ultrasound image analysis for first trimester assessment of fetal health. Physics in Medicine & Biology 64:185010. https://doi.org/10.1088/1361-6560/ab3ad1

40. van den Heuvel L.A. T, de Bruijn D, de Korte L. C, Ginneken B van (2018) Automated measurement of fetal head circumference using 2D ultrasound images. PLoS One 13:e0200412–e0200412. https://doi.org/10.1371/journal.pone.0200412

41. O'Shea RJ, Sharkey AR, Cook GJR, Goh V (2021) Systematic review of research design and reporting of imaging studies applying convolutional neural networks for radiological cancer diagnosis. Eur Radiol 31:7969–7983. https://doi.org/10.1007/s00330-021-07881-2

42. Lanowski J-S, Lanowski G, Schippert C, Drinkut K, Hillemanns P, Staboulidou I (2017) Ultrasound versus clinical examination to estimate fetal weight at term. Geburtshilfe Frauenheilkd 77:276–283

43. Mirghani HM, Weerasinghe S, Ezimokhai M, Smith JR (2005) Ultrasonic estimation of fetal weight at term: an evaluation of eight formulae. Journal of Obstetrics and Gynaecology Research 31:409–413

44. Barel O, Vaknin Z, Tovbin J, Herman A, Maymon R (2013) Assessment of the accuracy of multiple sonographic fetal weight estimation formulas: a 10-year experience from a single center. Journal of Ultrasound in Medicine 32:815–823

45. Faschingbauer F, Dammer U, Raabe E, Schneider M, Faschingbauer C, Schmid M, Schild RL, Beckmann MW, Kehl S, Mayr A (2015) Intrapartum sonographic weight estimation. Archives of Gynecology and Obstetrics 292:805–811. https://doi.org/10.1007/s00404-015-3720-3

Online supplementary material

MEDLINE and EMBASE (via OVID) search strategy

1. head ciurcumf$.mp.
2. f?etal biometry.mp.
3. head measur$.mp.
4. head biomet$.mp.
5. f?etal head.mp.
6. exp Artificial Intelligence/
7. exp Decision Making/
8. exp Deep Learning/
9. deep learning.mp.
10. exp Pattern Recognition, Automated/
11. automat$.mp.
12. cad.mp.
13. ai.mp.
14. Machine Learning/
15. cnn.mp.
16. convolutional neural network.mp
17. exp Image Interpretation, Computer-Assisted/ or exp Pattern Recognition, Automated/ or exp Image Processing, Computer-Assisted/ or exp Neural Networks, Computer/
18. 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17
19. exp Dimensional Measurement Accuracy/
20. accur$.mp.
21. exp “Reproducibility of Results”/
22. repro$.mp.
23. repeat$.mp.
24. feasib$.mp.
25. exp Feasibility Studies/
26. ultras$.mp.
27. exp Ultrasonography/
28. sono$.mp.

29. echo$.mp.

30. 19 or 20 or 21 or 22 or 23 or 24 or 25

31. 27 or 28 or 29 or 30

32. 1 or 2 or 3 or 4 or 5

33. 18 and 26 and 31 and 32

CINAHL (Via OBSCO) search strategy

S1. “head circumference”

S2. “head measurement”

S3. “f#etal biomet*”

S4. “f#etal biometry”

S5. “f#etal head”

S6. (MH “Head Circumference)

S7. (MH “Fetal Weight”) OR (MH “Fetal Monitoring+”)

S8. S1 OR S2 OR S3 OR S4 OR S5 OR S6 OR S7

S9. “artificial intelligence”

S10. “decision making”

S11. “deep learning”

S12. automat*

S13. cad

S14. ai

S15. cnn

S16. “convolutional neural network”

S17. “image interpre*”

S18. “computer vision”

S19. (MH “Artificial Intelligence”) OR (MH “Neural Networks (Computer)”)

S20. (MH “Computer Aided Design+”)

S21. (MH “Automation, Laboratory”) OR (MH “Automation+”)

S22. (MH “Image Processing, Computer Assisted+”) OR (MH “Radiographic Image Interpretation, Computer-Assisted+”) OR (MH “Image Interpretation, Computer Assisted+”)

S23. S9 OR S10 OR S11 OR S12 OR S13 OR S14 OR S15 OR S16 OR S17 OR S18 OR S19 OR S20 OR S21 OR S22

S24. accu*

S25. repeat*

S26. feasib*

S27. (MH “Reliability+”) OR (MH “Validity+”)

S28. S24 OR S25 OR S26 OR S27

S29. ultras*

S30. sono*

S31. (MH “Ultrasound Technologists”) OR (MH “Ultrasonography, Doppler, Transcranial”) OR (MH “Society of Radiologists in Ultrasound”) OR (MH “Ultrasonography, Prenatal+”)

S32. S29 OR S30 OR S31

S33. S8 AND S23 AND S28 AND S32

ArXiv and IEEE search strategy

1. "head cirumf*" OR "head measure*" OR biometry
2. automat* OR "convolutional neural network" OR "deep learning"
3. ultra* OR sono*
4. 1 and 2 and 3

QUADAS risk of bias checklist

	Patient selection[a,b]	Index test[d,e]	Reference Standard[g]	Flow and Timing[i,j]	Patient Selection[c]	Index test[f]	Reference standard[h]
Wu *et al* (2017)	High	Unclear	High	Low	Y	Y	Y
Sinclair *et al* (2018)	High	Low	High	Low	Y	Y	Y
Rong et *al* (2019)	High	Unclear	High	Low	Y	Y	Y
Oghli *et al* (2019)	High	Unclear	High	Low	Y	Y	Y
Sobhaninia *et al* (2019)	High	Low	High	Low	Y	Y	Y
Al-Bander *et al* (2020)	High	Unclear	High	Low	Y	Y	Y
Li *et al* (2020)	High	Low	High	Low	Y	Y	Y
Zhang *et al* (2020)	High	Low	High	Low	Y	Y	Y
Qiao *et al* (2020)	High	Low	High	Low	Y	Y	Y
Oghli *et al* (2021)	High	Unclear	High	Low	Y	Y	Y
Fiorentino *et al* (2021)	High	Low	High	Low	Y	Y	Y
Moccia *et al* (2021)	High	Low	High	Low	Y	Y	Y
Zeng *et al* (2021)	High	Low	High	Low	Y	Y	Y

[a] Was a consecutive or random sample of patients enrolled?

[b] Did the study avoid inappropriate exclusions?

[c] Are there concerns that the included patients and setting do not match the review question?

[d] Were the index test results interpreted without knowledge of the results of the reference standard?

[e] If a threshold was used, was it prespecified?

[f] Are there concerns that the index test, its conduct, or its interpretation differ from the review question?

[g] Is the reference standard likely to correctly classify the target condition?

[h] Are there concerns that the target condition as defined by the reference standard does not match the question?

[i] Did all patients receive the same reference standard?

[j] Were all patients included in the analysis?

Complete compliance checklist

Section / Topic	No.	Item													
TITLE / ABSTRACT			Cascaded fully convolutional networks for automatic prenatal ultrasound image segmentation	Human-level Performance On Automatic Head Biometrics In Fetal Ultrasound Using Fully Convolutional Neural Networks	Deriving external forces via convolutional neural networks for biomedical image segmentation	Automatic fetal biometry evaluation in ultrasound images using a deep learningbased approach	Fetal Ultrasound Image Segmentation for Measuring Biometric Parameters Using Multi Task Deep Learning	Improving Fetal Head Contour Detection by Object Localisation with Deep Learning	Automated measurement network for accurate segmentation and parameter modification in fetal head ultrasound images	A multiple-channel and atrous convolution network for ultrasound image segmentation	Dilated Squeeze-and-Excitation U-Net for Fetal Ultrasound Image Segmentation	Automatic fetal biometry prediction using a novel deep convolutional network architecture	A regression framework to head-circumference delineation from US fetal images	Mask-R2CNN: a distance-field regression version of Mask-RCNN for fetal-head delineation in ultrasound images	Fetal Ultrasound Image Segmentation for Automatic Head Circumference Biometry Using Deeply Supervised Attention-Gated V-Net
Author			Wu et al.[27]	Sinclair et al.[36]	Rong et al.[28]	Oghli et al.[31]	Sobhaninia et al.[32]	Al-Bander et al.[37]	Li et al.[29]	Zhang et al.[10]	Qiao et al.[38]	Oghli et al.[33]	Fiorentino et al.[34]	Moccia et al.[35]	Zeng et al.[30]
	1	Identification as a study of AI methodology, specifying the category of technology used (e.g., deep learning)	1	1	1	1	1	1	1	1	1	1	1	1	1
	2	Structured summary of study design, methods, results, and conclusions	1	1	1	0	1	1	1	1	1	1	1	1	1
INTRODUCTION															
	3	Scientific and clinical background, including the intended use and clinical role of the AI approach	1	1	1	0	1	1	1	1	1	1	1	1	1
	4	Study objectives and hypotheses	1	1	1	1	1	1	1	1	1	1	1	1	1
METHODS															
Study Design	5	Prospective or retrospective study	0	P	R	R	R	R	R	R	R	R	R	R	R
	6	Study goal, such as model creation, exploratory study, feasibility study, non-inferiority trial	1	1	1	1	1	1	1	1	1	1	1	1	1
Data	7a	Data sources	0	1	2	2	2	2	2	2	2	2	2	2	2
	7b	Different stages of pregnancy included	2	1	2	2	2	2	2	0	2	2	2	2	2
	7c	Appropriate inclusion of images with artifacts and/or difficult to obtain due to foetal position.	0	2	2	2	2	2	2	0	2	2	2	2	2
	7d	Appropriate inclusion of images in the later stages of pregnancy.	0	1	1	1	1	1	1	0	1	1	1	1	1
	7e	BMI of patients	0	0	0	0	0	0	0	0	0	0	0	0	0
	7f	Multiple ultrasound machine vendors	0	1	1	1	1	1	1	0	1	1	1	1	1
	8	Eligibility criteria: how, where, and when potentially eligible participants or studies were identified (e.g., symptoms, results from previous tests, inclusion in registry, patient-care setting, location, dates)	0	0	1	1	1	1	1	0	1	1	1	1	1
	9	Data pre-processing steps	1	1	1	0	1	1	1	1	1	1	1	1	1
	10	De-identification methods	0	1	1	1	1	1	1	0	1	1	1	1	1
Ground Truth	11	Definition of ground truth reference standard, in sufficient detail to allow replication	1	1	1	1	1	1	1	0	1	1	1	1	1
	12	Source of ground-truth annotations; qualifications and preparation of annotators	1	1	1	1	1	1	1	0	1	1	1	1	1
Data Partitions	13a	Intended sample size	1	1	1	1	1	1	1	1	1	1	1	1	1
	13b	Rational for sample size	0	0	0	0	0	0	0	0	0	0	0	0	0
	13c	How data were assigned to partitions	0	1	1	1	1	1	1	1	1	1	1	1	1
	13d	Specify proportions	0	0	0	0	0	0	0	0	0	0	0	0	0
	14	Level at which partitions are disjoint (e.g., image, study, patient, institution)	1	1	1	1	1	1	1	0	1	1	1	1	1
Model	15	Detailed description of model, including inputs, outputs, all intermediate layers and connections	2	2	2	1	2	2	2	2	2	1	2	2	2
	16	Software libraries, frameworks, and packages	0	0	1	0	1	1	0	1	1	1	1	1	1
	17	Initialization of model parameters (e.g., randomization, transfer learning)	1	2	2	0	2	2	2	2	2	1	2	2	2
Training	18a	Details of training approach, including data augmentation, hyperparameters, number of models trained	1	1	1	0	2	1	2	2	2	1	2	2	2
	18b	Distinction between training and validation	0	1	0	0	1	0	0	0	1	0	1	1	1
Evaluation	19	Metrics of model performance	1	1	1	1	1	1	1	1	1	1	1	1	1
	20	Statistical measures of significance and uncertainty (e.g., confidence intervals)	0	0	0	1	1	1	1	1	1	1	1	1	1
	21	Robustness or sensitivity analysis	0	0	0	0	0	0	1	0	0	0	1	1	1
	22	Validation or testing on external data	1	1	1	1	1	1	1	1	1	1	1	1	1
RESULTS						0									
Data	23	Flow of participants or cases, using a diagram to indicate inclusion and exclusion	0	0	0	0	0	0	0	0	0	0	0	0	0
	24	Demographic and clinical characteristics of cases in each partition	0	0	0	0	0	0	0	0	0	0	0	0	0
Model performance	25	Performance metrics for optimal model(s) on all data partitions	1	1	1	1	1	1	1	1	1	1	1	1	1
	26	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	1	1	1	1	1	1	1	1	1	1	1	1	1
DISCUSSION															
	27	Study limitations, including potential bias, statistical uncertainty, and generalizability	0	0	0	0	0	0	1	0	1	0	1	1	1
	28	Implications for practice, including the intended use and/or clinical role	1	1	1	1	1	1	1	1	1	1	1	1	1