

# ATLAS: Development and Evaluation of an AI - Enabled Framework for Anatomically Traceable Measurement Assurance in Fetal-Head Ultrasound Biometry

Dian Tjondronegoro

[d.tjondronegoro@griffith.edu.au](mailto:d.tjondronegoro@griffith.edu.au)

Griffith University

Elizabeth Irenne Yuwono

Griffith University

Fabricio da Silva Costa

Gold Coast University Hospital

---

## Research Article

**Keywords:** medical informatics, fetal ultrasound, fetal-head biometry, measurement validity, patient-exclusive validation, data provenance, weak supervision, external validation, edge computing, ISUOG alignment

**Posted Date:** October 12th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-10833481/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# ATLAS: Development and Evaluation of an AI -Enabled Framework for Anatomically Traceable Measurement Assurance in Fetal-Head Ultrasound Biometry

Dian Tjondronegoro<sup>1</sup>, Elizabeth I. Yuwono<sup>1</sup>, Fabricio da Silva Costa<sup>2</sup>

<sup>1</sup> Griffith Business School, Griffith University, Brisbane, Australia

<sup>2</sup> Maternal-Fetal Medicine Unit, Gold Coast University Hospital and Griffith University, Gold Coast, Australia

*Corresponding author: Dian Tjondronegoro (d.tjondronegoro@griffith.edu.au), Elizabeth I. Yuwono (e.yuwono@griffith.edu.au)*

## Abstract

**Background:** Fetal-head biometry is important for estimating gestational age, monitoring fetal development, and identifying abnormalities. As AI-enabled ultrasound systems become more prevalent, automated measurements may appear credible even with mislocalization, suboptimal imaging planes, or absent anatomical landmarks. Clinicians may receive measurements without adequate evidence on their origin or reliability. ATLAS is designed as an evidence-traceable framework to establish conditions for valid measurements.

**Methods:** ATLAS consists of a data integrity layer, Find, Confirm, and Measure. The integrity layer performs dataset-agnostic checks for image validity, label consistency, and patient-exclusive partitioning. Find localizes the fetal cranial region; Confirm verifies the anatomical structures required for valid biometry. Measure calculates biparietal diameter, occipitofrontal diameter, and head circumference using two methods: a deterministic localization geometry and a learned segmentation. Training used 3,790 eligible frames from 1,840 patients from public hospital-derived HC18/FETAL\_PLANES\_DB resources, with patient-exclusive training, validation, and testing; external evaluation used FP (880 frames/460 subjects) and UCL (55/15) cohorts. Outcomes were localization, CSP/LV evidence, and biometric error. Performance measures included availability/IoU, F1, mean absolute error, signed error, and prespecified joint BPD/HC tolerance. Site-specific adaptation and data-free parameter aggregation were added as secondary technical analyses for future edge deployment.

**Results:** Under controlled data-fault tests, the data integrity layer detected every introduced defect. Internal validation demonstrated that Find localized all frames (mean IoU 0.962); Confirm F1 was 0.894 for CSP and 0.819 for LV. Deterministic geometry met both biometric tolerances in 91.6% frames versus 84.1% for segmentation, reducing BPD MAE by 0.291 mm (95% CI 0.225–0.355) and improving joint-tolerance performance by 7.5pp (95% CI 5.2–9.9), meaning measurements following positive Confirm being more accurate. Externally, Find localized 99.9% (FP) and 100% (UCL) of frames, with deterministic geometry measurement performing better on FP (91.2%) and segmentation on UCL (79.6%). Site-specific adaptation improved localization but reduced biometry accuracy, which aggregation partially restored. The most memory-intensive ATLAS pathway required 16.50 MB and 57.32 ms per frame.

**Conclusions:** This study presents a compact, evidence-traceable framework that treats operationalizing selected ISUOG-relevant anatomical and biometric requirements as computational quality-control states. The system achieves performance comparable to existing approaches, while prioritizing measurement validity and transparency as core design principles. These findings underscore the necessity for clinical AI systems to provide not only accurate results but also interpretable and auditable evidence to support trustworthy integration.

**Keywords:** medical informatics; fetal ultrasound; fetal-head biometry; measurement validity; patient-exclusive validation; data provenance; weak supervision; external validation; edge computing; ISUOG alignment

## Background

Fetal-head biometry is a routine yet critical component of antenatal ultrasound. Measurements such as biparietal diameter (BPD), occipito-frontal diameter (OFD), and head circumference (HC) are essential for pregnancy dating, fetal-growth surveillance, and the assessment of abnormal head size. These metrics inform downstream calculations of estimated fetal weight, which influence decisions regarding repeat scanning, referral, and ongoing surveillance of mother and fetal health (1). These values are not interpreted in isolation. Valid measurements require that the image represents the

50 correct, symmetrical cranial plane, the expected internal anatomy is present, the measurement  
51 geometry is appropriate, and pixel spacing allows reliable conversion from image pixels to  
52 millimetres (2) As artificial intelligence (AI)-enabled ultrasound systems increasingly developed to  
53 support or automate plane recognition and biometric measurement, the same conditions remain  
54 necessary for interpreting the resulting values. If any of these conditions are not met, the AI-enabled  
55 automated value may appear numerically plausible while lacking adequate support from the  
56 underlying image.

57 This gap is critical because, in obstetric ultrasound, measuring an inappropriate imaging plane is a  
58 major source of biometric error, often exceeding the variability introduced by caliper placement on  
59 an otherwise correct image (3). An oblique or incorrectly positioned section can shift a measurement  
60 by several millimetres, leading to systematic misestimation of fetal size, distortion of growth  
61 centiles, and subsequent clinical decisions (4). A recent systematic review has identified that a  
62 precise value generated by an AI-enabled system is not necessarily accurate (5). Automated systems  
63 may exacerbate this risk when outputs are produced despite unsuitable acquisition conditions or  
64 when AI confidence scores do not expose the specific anatomical or measurement conditions  
65 underlying the result, allowing failure to remain difficult to trace (6). Under time constraints, a well-  
66 drawn ellipse and a confident value may prompt user acceptance without independent verification,  
67 a phenomenon documented as automation bias in the literature (7). End-to-end systems further limit  
68 the clinician's ability to review, question, or reject results due to the lack of accessible evidence (8).  
69 The resulting risk reflects an interaction between model behavior, the information presented to  
70 users, and human reliance on automation.

71 The persistence of this problem spans multiple disciplines. It encompasses anatomical challenges,  
72 such as the difficulty of reliably capturing a standard plane using freehand two-dimensional  
73 ultrasound (9); machine-learning limitations, where models optimized for average accuracy on  
74 curated data learn shortcut correlations and remain confident outside the training distribution (5,7);

75 human-factors issues, as a single numerical output provides users with no basis for critical  
76 evaluation (4); standardization barriers, since narrative guidelines are difficult to translate into  
77 machine-readable constraints (1,10); data limitations, with public datasets under-representing  
78 pathological and off-plane cases and frame-level splitting inflating apparent generalization (11,12);  
79 and deployment constraints (13,14), as the most robust models are often too resource-intensive for  
80 point-of-care settings. These implementation constraints may be particularly relevant in rural,  
81 remote, and resource-limited settings where access to specialist obstetric-ultrasound expertise and  
82 advanced computational infrastructure is more constrained (15). Trustworthiness depends on this  
83 interconnected chain of conditions and is only as strong as its weakest link. Improving any single  
84 component in isolation does not resolve the overall issue.

85 Most fetal-ultrasound AI research has advanced individual components of this chain, including  
86 standard-plane classification, head segmentation, landmark detection, direct biometry, and whole-  
87 examination automation (16–18). Commercial, regulatory-cleared ultrasound systems also provide  
88 real-time plane recognition and for some, automated biometry during image acquisition (19). These  
89 acquisition-facing capabilities are clinically important, but they do not remove the need for  
90 subsequent quality assurance. International Society of Ultrasound in Obstetrics and Gynecology  
91 (ISUOG) Practice Guidelines recommend routine quality-control of fetal biometric images,  
92 including image storage, review and assessment of measurement reproducibility, while  
93 standardized fetal-ultrasound guidance requires an image to document each biometric measurement  
94 (1,20). This creates an informatics requirement beyond real-time plane recognition: the evidence  
95 needed to review measurement quality should remain identifiable and linked to the measurement  
96 after acquisition.

97 This work introduces ATLAS (Anatomically-Traceable Localization, Assessment and  
98 Segmentation for Fetal Neurosonography), as an evidence-traceability framework intended to  
99 complement, rather than replace, acquisition-facing systems by supporting post-acquisition review

100 and quality assurance of fetal-head biometry by sonographers, reporting clinicians, and clinical  
101 quality or governance teams. ATLAS incorporates a patient-audited data-integrity layer and three  
102 interdependent stages that reflect clinical reasoning: Find, which localizes the cranial field and  
103 determines measurement eligibility; Confirm, which assesses the cavum septi pellucidi (CSP) and  
104 lateral ventricles (LV) prior to measurement; and Measure, which performs segmentation-derived  
105 biometry. All stages are linked through an anatomically-traceable provenance ledger. The central  
106 design principle inverts the conventional approach: instead of generating a value and subsequently  
107 explaining it, ATLAS first verifies that a measurement is supported, then derives a value that  
108 remains linked to its eligibility, anatomical evidence, pixel-spacing source, and output-availability  
109 state. This approach reframes automated biometry as a measurement-assurance problem: rather than  
110 estimating a value alone, ATLAS retains evidence about the conditions under which that value was  
111 generated.

112 Its proposed contribution is to retain a structured record linking each reported measurement to  
113 cranial localization, selected anatomical evidence relevant to plane adequacy, measurement  
114 geometry, pixel-scale provenance, and output availability. This record could support subsequent  
115 quality review and audit by making the basis of an automated measurement retrievable rather than  
116 leaving only the final numerical value. Accordingly, this study aimed to develop and retrospectively  
117 evaluate ATLAS for evidence-traceable fetal-head biometry, assessing cranial localization, selected  
118 anatomical evidence, and BPD/OFD/HC measurement under patient-exclusive internal testing and  
119 independent external evaluation. A secondary objective was to assess how site-specific adaptation  
120 and data-free parameter aggregation affected cross-site localization and biometric performance, to  
121 inform future investigation of cross-institutional model adaptation where raw imaging data cannot  
122 be centrally pooled.

123

## 124 **Methods**

### 125 **Reporting framework, study design, and intended use**

126 Reporting followed TRIPOD+AI, and APPRAISE-AI guidelines for high-quality study, adapted  
127 where appropriate for a retrospective methodological development and benchmark study (21–24).  
128 Complete checklists for these guidelines are available in Additional Files 1 to 3. The study involved  
129 patient-exclusive internal and external evaluations, including patient-exclusive development,  
130 locked internal validation, and external evaluation. For internal training and validation, the fetal-  
131 head annotation resource from HC18 and FETAL\_PLANES\_DB was used, with data divided into  
132 controlled, leakage-proof patient-exclusive partitions. External evaluation utilized independent FP  
133 and UCL cohorts from the FetalBiometry-MultiCentre-Landmarks resource to assess whether  
134 localization and biometric performance could be generalized beyond the training data source.

135 The study analyzed previously acquired ultrasound frames to assess technical performance, failure  
136 detection capability, and cross-cohort robustness in a retrospective context. The evaluation did not  
137 include real-time scanning, autonomous diagnosis, clinical decision-making, patient outcomes, or  
138 application of the method on an ultrasound device. Therefore, the results should be interpreted as  
139 preliminary methodological evidence rather than evidence of clinical effectiveness. Prospective  
140 validation during routine ultrasound practice is required before clinical implementation.

### 141 **Clinical reference framework and scope of computational operationalization**

142 The design of ATLAS was informed by the ISUOG Practice Guidelines for ultrasound assessment  
143 of fetal biometry and growth (20). ISUOG emphasizes that fetal measurements require clear images  
144 with sufficient magnification and correct depiction of relevant landmarks to permit accurate  
145 measurement placement, together with regular quality control (Figure 1). For cephalic biometry  
146 specifically, the objective image-quality framework comprises a symmetrical cranial plane,  
147 visualization of the thalami and cavum septi pellucidi (CSP), absence of the cerebellum, sufficient  
148 magnification such that the fetal head occupies more than half of the image, and correct placement  
149 of the calipers and circumference ellipse. These criteria were treated as complementary dimensions

150 of image and measurement quality rather than as interchangeable indicators of an acceptable  
151 biometric plane.

152 ATLAS was designed to translate this clinical reasoning into a structured computational pathway  
153 while preserving the distinction between what was directly evaluated, what was represented but not  
154 clinically adjudicated, and what remained outside the available benchmark supervision. The  
155 architecture therefore did not attempt to reduce image adequacy to a single detected landmark or  
156 model confidence. Instead, different stages represented different components of the evidence  
157 required to interpret a biometric result: the data-integrity layer retained source and scale  
158 provenance; Find established whether a fetal cranial region could be localized and recorded the  
159 geometry of that localization; Confirm assessed selected internal anatomical evidence; Measure  
160 exposed the geometry and scale used to derive BPD, OFD, and HC; and the structured output  
161 retained these components alongside measurement availability.

162 The extent of clinical-guideline operationalization was constrained by the annotations available in  
163 the open datasets. HC18 provides expert fetal-head annotations and image-scale information that  
164 support evaluation of cranial localization and biometric geometry, but it does not provide  
165 independent image-level labels for all ISUOG cephalic quality criteria. In particular, expert labels  
166 were not available for thalamic visualization, cerebellar exclusion, overall plane symmetry, or  
167 complete image-level adequacy. Accordingly, these criteria could not be used as supervised targets  
168 or as complete image-level ground truth in the present benchmark. The ISUOG criteria therefore  
169 served as the clinical reference framework for system design, interpretation, and scope, rather than  
170 as a fully annotated reference standard against which every HC18 frame could be classified as  
171 clinically adequate or inadequate.

172 Within the anatomical component of this framework, CSP was selected as a feasible exemplar of  
173 small-structure evidence assessment because it was both clinically relevant to the ISUOG cephalic  
174 quality framework and explicitly annotated in the available annotation resource, whereas the

175 thalami, although visually relevant to the trans-thalamic plane, were not provided as independently  
176 localized training targets in that resource (25). Detailed Dataset provided in the later subsection  
177 (Datasets and development-evaluation design). CSP is one component of the ISUOG cephalic  
178 image-quality framework in addition to thalami and flax. CSP detection alone is not sufficient to  
179 establish an acceptable biometric plane. The lateral ventricle (LV) was retained as a second  
180 anatomical target to evaluate whether the same evidence architecture could represent more than one  
181 intracranial structure. Although LV visualization is clinically informative in fetal-brain assessment,  
182 it is not one of the cephalic biometric-image quality criteria specified in the 2019 ISUOG objective  
183 quality framework. Confirm therefore operationalized selected anatomical evidence that was both  
184 clinically relevant and available for supervised computational assessment, rather than complete  
185 standard-plane verification.

186 Other ISUOG dimensions informed the ATLAS' staged Find–Confirm–Measure design. Find  
187 retains the location and extent of the fetal-head region relative to the original ultrasound frame,  
188 thereby preserving information relevant to image positioning and magnification for subsequent  
189 review. This design relates particularly to the ISUOG requirement for sufficient magnification of  
190 the fetal head, while retention of the original full-frame context allows the reviewer to inspect  
191 broader acquisition characteristics that cannot be inferred from the cropped region alone. Find does  
192 not, however, apply an expert-validated threshold to classify magnification, centering, or plane  
193 symmetry as clinically adequate; these remain reviewable image properties by clinical users rather  
194 than automated acceptance criteria.

195 Confirm addresses the landmark-evidence component of the framework by making selected internal  
196 anatomy explicit within the localized cranial region. CSP was operationalized as a clinically  
197 relevant small-anatomy exemplar, with LV retained as supplementary intracranial evidence, while  
198 thalamic visualization and cerebellar exclusion were not independently automated in the present

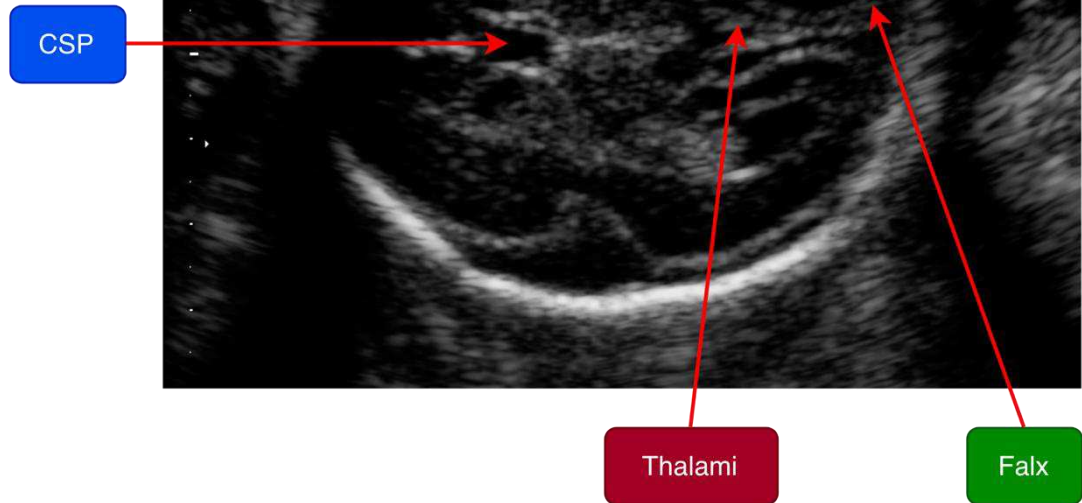
199 implementation. Consequently, the Confirm output provides evidence relevant to plane  
200 interpretation rather than a complete determination of standard-plane adequacy.

201 Measure addresses the measurement-execution component of the framework by retaining the  
202 predicted fetal-head contour or Find-box geometry, the fitted ellipse, the pixel-spacing source, and  
203 the resulting BPD, OFD, and HC. These BPD, OFD, and HC measurements follow the ISUOG fetal  
204 biometry guideline, ensuring that the expert reference measurements and AI-derived measurements  
205 use the same caliper and ellipse conventions. The visible geometry allows inspection of the basis  
206 for the automated measurement, but ATLAS does not replace expert judgment on whether the  
207 measurement placement is clinically acceptable or consistent with the locally adopted biometric  
208 methodology.

209 Accordingly, criteria that were not independently automated, including overall plane symmetry,  
210 thalamic visualization, cerebellar exclusion, and complete clinical assessment of magnification and  
211 measurement placement, remain within the clinician's domain of interpretation. ATLAS supports  
212 this assessment by retaining the original acquisition context, a focused anatomical-evidence view,  
213 explicit measurement geometry, scale provenance, and stage-specific outputs in a measurement-  
214 level record. These features make the available evidence accessible for review, but they were not  
215 evaluated as substitutes for expert assessment of complete ISUOG image adequacy. Because  
216 individual quality criteria are not interchangeable, ATLAS retains every evidence as separate states  
217 rather than collapsing them into a single automated judgement of plane adequacy.

218 This architecture is also consistent with the quality-assurance context of fetal biometry. ISUOG  
219 recommends routine quality-control checks of biometric images, including expert image review and  
220 reproducibility assessment of stored images. ATLAS was therefore designed to retain  
221 measurement-level evidence in a form that could support subsequent review and audit.

Patient01126\_  
Plane3\_4\_of\_5



222

223 Figure 1. ISUOG-aligned anatomical reference for the fetal head biometry plane on an internal test  
224 image

## 225 **System overview**

226 ATLAS consists of a data-integrity layer and three analysis stages: Find, Confirm, and Measure,  
227 followed by generation of a structured measurement-level audit record (Figure 2).

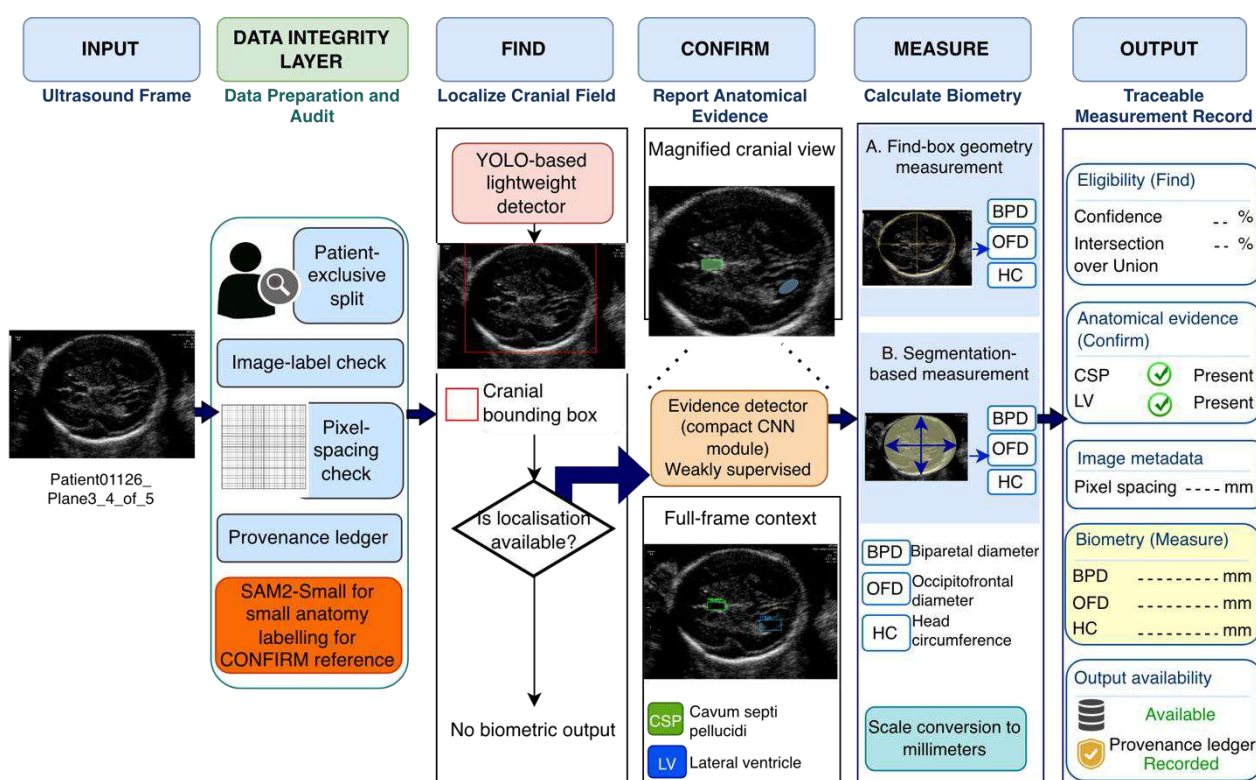
228 The data-integrity layer partitions the source ultrasound data into patient-exclusive development  
229 and evaluation sets and generates the labels required for each stage. Additionally, it preserves the  
230 origin of each annotation, patient and image identifiers, and pixel-spacing information. Its purpose  
231 was to ensure that the evidence used by subsequent stages remained attributable to its source and  
232 that computationally derived labels were not treated as equivalent to independent expert  
233 annotations.

234 The Find stage identifies the fetal cranial region within the complete ultrasound frame. Its primary  
235 objective is to localize the fetal cranial region and establish whether a measurement-generating  
236 region is available. If no region is detected with sufficient confidence, the system records the  
237 measurement as unavailable, rather than deriving biometry from an uncertain image region.

238 The Confirm stage examines a focused view of the localized cranial region for selected internal  
 239 anatomical evidence. It assessed CSP and LV visibility, with CSP representing one clinically  
 240 relevant component of the ISUOG cephalic image-quality framework and LV providing  
 241 supplementary intracranial evidence.

242 The Measure stage estimates the geometry of the fetal head. It converts these measurements into  
 243 BPD, OFD, and HC using the recorded pixel spacing. BPD measures the transverse width of the  
 244 head, OFD measures the anteroposterior dimension, and HC reflects overall head size. These  
 245 parameters are routinely used to support pregnancy dating and fetal growth assessment.

246 The Find, Confirm, and Measure stages are deliberately separated to enable explicit examination of  
 247 the basis for each biometric result. The resulting record documents whether the fetal head was  
 248 localized, which anatomical structures were detected, how head geometry was estimated, and how  
 249 image dimensions were converted into clinical units. Consequently, ATLAS does not treat the final  
 250 biometric value as self-sufficient; instead, it links this value to the localization, anatomical evidence,  
 251 measurement geometry, and scale information that support its interpretation.



253 Figure 2. ATLAS system pipeline

254 **Datasets and experimental design**

255 Table 1 provides an overview of the internal development partitions and external evaluation cohorts  
256 at both the patient and image-frame levels.

257 ATLAS development used the open large-scale fetal-head annotation resource, which consolidates  
258 and re-annotates fetal ultrasound images originating from the HC18 and FETAL\_PLANES DB  
259 datasets (25,26). The published resource contains 3,832 fetal-head ultrasound images and provides  
260 annotations of the fetal brain, CSP, and LV, together with information required for image-level  
261 analysis. The brain was annotated using an elliptical region, whereas CSP and LV were represented  
262 using oriented rectangular annotations. The published annotation process included clinical  
263 validation involving a senior attending physician and a radiologic technologist (24).

264 The first source, HC18, is a fetal-head circumference dataset acquired at Radboud University  
265 Medical Center, Nijmegen, the Netherlands collected between May 2014 to May 2015 during  
266 routine ultrasound screening (26). Inclusion criteria specified pregnant women undergoing routine  
267 ultrasound screening with fetuses exhibiting no growth abnormalities. The original HC18 cohort  
268 contains 1,334 two-dimensional ultrasound images acquired from 551 pregnant women using GE  
269 Voluson E8 and Voluson 730 systems and was released as 999 training images and 335 independent  
270 test images. The images have a native resolution of  $800 \times 540$  pixels, with image-specific pixel  
271 sizes reflecting acquisition depth and zoom settings. For the derived annotation resource, the  
272 publicly annotated HC18 images provided fetal-head geometry and physical image-scale  
273 information relevant to cranial localization and biometric measurement (25).

274 The second database, FETAL\_PLANES\_DB, comprises a large collection of ultrasound images  
275 obtained routinely at two hospitals (the Hospital Clinic and the Hospital Sant Joan de Deu in  
276 Barcelona, Spain) between October 2018 and April 2019 as part of standard clinical practice

277 (27). Images were acquired by various operators using different ultrasound machines. Inclusion  
278 criteria required pregnant women attending routine second and third trimester screenings, excluding  
279 cases with multiple pregnancies, congenital malformations, or aneuploidies. The original database  
280 contains 12,400 images from 1,792 patients, categorized into six main anatomical groups: fetal  
281 abdomen, brain, femur and thorax, maternal cervix, and other less common planes. Fetal-brain  
282 images are further subdivided into trans-thalamic, trans-cerebellar, and trans-ventricular planes.  
283 The dataset also includes metadata on patient number, ultrasound machine, and operator. Only fetal-  
284 head images included in the published fetal-head annotation resource were used for ATLAS  
285 development; other anatomical categories from FETAL\_PLANES\_DB were excluded (25).

286 The published fetal-head annotation resource contains 3,832 image records (25). During source-  
287 image and annotation reconciliation for the present study, 3,790 frames could be linked to the  
288 required image and provenance records and were retained for the study; the remaining 42 published  
289 annotation records were therefore excluded before partitioning. The final analysis-ready cohort is  
290 specified in Table 1.

291 Patient identifiers encoded in, or recoverable from, the source records were then used to assign all  
292 retained images belonging to the same patient to a single mutually exclusive partition: training,  
293 validation, or locked testing. Automated integrity checks (Data Integrity Layer Subsection) halted  
294 the data-processing workflow if the same patient identifier was detected in more than one partition.  
295 The resulting cohort comprised 1,840 patients and 3,790 image frames, allocated approximately  
296 70:15:15 across training, validation, and locked-test partitions. The training partition was used to  
297 estimate model parameters, whereas the validation partition informed model selection and other  
298 prespecified training decisions. The locked-test partition remained isolated until final evaluation.  
299 No image belonging to a locked-test patient was used for model training, checkpoint selection,  
300 threshold selection, pseudo-reference generation, or curriculum design.

301 For physical-scale recovery, image-specific pixel spacing was used when available in the source  
302 metadata and within the prespecified plausible range of 0.04–0.40 mm per pixel. Pixel spacing  
303 specifies the physical distance represented by one image pixel and is required to convert image-  
304 domain measurements into physical units. ATLAS first estimates the relevant cranial geometry in  
305 pixels and subsequently applies the corresponding image scale to obtain BPD and OFD in  
306 millimeters and to derive HC from the resulting physical head geometry. The provenance of the  
307 scale value was retained for each output so that the measurement pathway remained auditable.

308 When image-specific spacing was missing or outside the prespecified plausible range, a fixed  
309 fallback of 0.125 mm per pixel was applied. This value was calculated as the median of valid pixel-  
310 spacing values in the training partition only. It was established without reference to validation or  
311 test data and remained unchanged throughout evaluation to maintain the non-leakage data principle.  
312 Each output record indicated whether its measurement used image-specific metadata or the fallback  
313 value, enabling results based on imputed scale information to be distinguished during audit and  
314 interpretation.

315 External evaluation used the Fetal Planes (FP) and University College London (UCL) Hospital  
316 sources from the FetalBiometry-MultiCentre-Landmarks resource (28). FP contains 3,090 image  
317 frames from 1,047 patients at Vall d'Hebron and Sant Joan de Déu, Barcelona, Spain. UCL includes  
318 424 image frames from 51 patients at UCL Hospital, United Kingdom. Both sources contain head,  
319 abdomen, and femur anatomical planes; however, only the head plane images were used in this  
320 study. Both datasets were collected using GE Voluson systems and annotated using manual  
321 landmarks. These datasets were suitable for external evaluation because they represent independent  
322 fetal-ultrasound cohorts outside HC18 and provide subject identifiers, expert anatomical landmarks,  
323 image dimensions, and pixel-spacing information necessary to evaluate cranial localization and  
324 derive biometric measurements. For each external source, the predefined subject-exclusive test

partition was used for evaluation, ensuring that no subject appeared in more than one development or test partition.

Reference head contours, external landmarks, and reference biometric values were not provided to ATLAS during internal or external inference. These data were used only after inference to evaluate localization and biometric performance. Detailed training and testing hyperparameters are provided in Additional File 4.

Table 1. Patient-exclusive source and external partitions

| Partition           | Patients/subjects      | Frames                 | Pixel-spacing status                            |
|---------------------|------------------------|------------------------|-------------------------------------------------|
| Training            | 1,290 of 1,840 (70.1%) | 2,629 of 3,790 (69.4%) | 2,352 metadata (89.5%);<br>277 fallback (10.5%) |
| Internal validation | 271 of 1,840 (14.7%)   | 575 of 3,790 (15.2%)   | 512 metadata (89.0%);<br>63 fallback (11.0%)    |
| Internal test       | 279 of 1,840 (15.2%)   | 586 of 3,790 (15.5%)   | 522 metadata (89.1%);<br>64 fallback (10.9%)    |
| External FP test    | 460 of 909 (50.6%)     | 880 of 1,637 (53.8%)   | 880/880 metadata (100%);<br>0 fallback          |
| External UCL test   | 15 of 47 (31.9%)       | 55 of 159 (34.6%)      | 35 metadata (63.6%);<br>20 fallback (36.4%)     |

Note: Metadata indicates that pixel spacing was obtained from the dataset metadata for the corresponding image. Fallback indicates that valid image-specific spacing was unavailable and the prespecified HC18 training-set median of 0.125 mm per pixel was used.

### Data integrity layer

Public fetal-ultrasound datasets are typically developed for specific research tasks rather than as comprehensive training resources for multi-stage measurement systems. For instance, datasets may include image-level standard-plane classifications, fetal-head contours, or biometric landmarks, but often lack the complete set of patient identifiers, internal-anatomy annotations, measurement geometry, and physical image-scale information required by ATLAS (25,29,30). The training resources provide patient-linked ultrasound frames, expert fetal-head annotations, and pixel-spacing information suitable for cranial localization and biometry. However, its expert reference standard does not include pixel-level masks for the CSP and LV. The data-integrity layer was designed not to fabricate missing clinical ground truth, but to convert available source evidence into clearly defined, stage-specific resources while preserving the provenance and limitations of each derived label.

347 The data-integrity layer initially converted distributed source files into a unified frame-level data  
348 structure. It resolved nested archives and source directories, verified the availability of referenced  
349 images, removed duplicate entries, standardized patient and image identifiers, ensured each image  
350 was matched to the expected annotation, and retained empty annotation files as explicit negative  
351 examples. Retaining explicit negatives was essential to distinguish a confirmed absence of an  
352 annotated structure from a missing or unprocessed label. The pipeline also validated image  
353 dimensions and pixel-spacing information, recording each frame in a common audit ledger. When  
354 a critical inconsistency was detected, such as patient overlap between partitions, a missing source  
355 image, or an image-annotation mismatch, dataset construction was halted and the discrepancy  
356 required review. Records were not silently removed, reassigned, or treated as valid training  
357 examples.

358 For the Find stage, expert fetal-head ellipse annotations were converted into axis-aligned bounding  
359 boxes enclosing the annotated head. These bounding boxes provided a consistent target for object-  
360 detection training while maintaining a direct link to the original expert head annotation. The source  
361 patient identifier, image identifier, annotation coordinates, and pixel-spacing record were preserved  
362 to ensure that every predicted cranial region could be traced back to its reference evidence.

363 For Confirm, human-curated CSP/LV bounding boxes associated with the HC18 source images  
364 were obtained from the auxiliary anatomical annotation resource and linked to the corresponding  
365 patient and image records (25). These boxes were then remapped from the full ultrasound frame  
366 into the localized cranial coordinate system. The original HC18 reference standard did not provide  
367 expert pixel-level CSP/LV masks. Accordingly, the boxes represented the available human-curated  
368 anatomical evidence.

369 Because the source dataset did not provide expert-drawn pixel-level outlines of these structures,  
370 Segment Anything Model 2–Small (SAM2-Small) was employed during dataset preparation  
371 (31,32). SAM2-Small is a general-purpose, prompt-guided image-segmentation model that, when

372 supplied with a box indicating the expected location of a structure, estimates which pixels within  
373 that region are most likely to belong to the structure. Each curated CSP or LV box was provided to  
374 SAM2-Small as a spatial prompt, and the model generated a provisional mask outlining the likely  
375 structure. This process was performed offline during training data construction. SAM2-Small was  
376 not used during subsequent ATLAS analysis of ultrasound images and did not generate the final  
377 fetal-head measurements.

378 The resulting masks served as approximate training labels for the lightweight Confirm model. These  
379 masks enabled Confirm to learn a more detailed representation of CSP and LV locations than would  
380 be possible from boxes alone, but they were not treated as expert annotations or independent clinical  
381 ground truth. Each generated mask remained linked to its originating human-curated box, source  
382 image, patient identifier, and transformation record.

383 For the Measure stage, each fetal-head box was converted into an ellipse represented by 72 equally  
384 spaced boundary points. These points corresponded to five-degree intervals around the ellipse,  
385 producing a smooth and reproducible closed contour suitable for training a segmentation model.  
386 The 72-point representation was an engineering choice to approximate fetal-head geometry  
387 consistently; it was not intended as a clinical annotation standard or as an independently drawn  
388 expert contour. Although both Confirm and Measure involve segmentation, they serve distinct purposes  
389 and utilize different types of training targets. SAM2-Small was used exclusively during dataset  
390 preparation to generate provisional masks for the smaller internal structures assessed by Confirm  
391 and was not used during inference. In contrast, the Measure segmentation model functions as an  
392 active component of ATLAS during inference. It predicts the overall fetal-head contour, fits an  
393 ellipse to that contour, and uses the fitted geometry and recorded pixel spacing to calculate BPD,  
394 OFD, and HC. Thus, Confirm evaluates anatomical evidence within the head, whereas Measure  
395 estimates the external head geometry from which biometric measurements are derived.

396 Segmentation was incorporated into Measure because it produces a visible head contour and fitted  
397 ellipse that can be reviewed alongside the resulting biometric values. This approach also enabled  
398 evaluation of whether learning an explicit representation of fetal-head shape improved  
399 measurement accuracy compared with calculating BPD, OFD, and HC directly from the rectangular  
400 region identified by Find. Since the Measure training contours were derived from cranial-box  
401 geometry rather than independently drawn expert head contours, the simpler Find-box measurement  
402 pathway was retained as a prespecified comparator. This comparison allowed the contribution of  
403 learned shape estimation to be evaluated empirically.

404 Every source annotation and derived label remained linked to the relevant patient, image, original  
405 annotation, transformation method, pixel-spacing source, and processing version. This provenance  
406 structure clarifies which labels originated from experts, which were human-curated boxes, and  
407 which were generated computationally, thereby preventing these different forms of evidence from  
408 being interpreted as equivalent reference standards.

#### 409 **Image preprocessing**

410 All stages employed a consistent, prespecified image-preprocessing pipeline. Ultrasound images  
411 frequently exhibit speckle, a grainy pattern resulting from the interaction of reflected sound waves,  
412 which can obscure subtle anatomical details (33). Bilateral filtering was applied to reduce this grain-  
413 like variation by smoothing neighboring pixels with similar intensity while preserving sharp  
414 intensity transitions (34). This approach suppressed noise without blurring important anatomical  
415 boundaries.

416 Contrast-limited adaptive histogram equalization was subsequently applied to enhance local  
417 contrast (35,36). Instead of uniformly adjusting brightness across the entire image, this technique  
418 increases contrast within small regions, ensuring that structures remain visible in both dark and  
419 bright areas. The contrast enhancement is constrained to prevent excessive amplification of noise.

420 The preprocessing procedure and all parameter settings were established prior to evaluation and  
421 applied consistently throughout model development, validation, and testing. Detailed filter  
422 parameters, image augmentation settings, software versions, and checkpoint hashes are provided in  
423 Additional file 1.

#### 424 **Find stage for cranial-field localization**

425 The Find module was developed to identify the fetal cranial region within the complete ultrasound  
426 frame and to assess whether localization confidence was sufficient for biometric analysis. An  
427 image-based object-detection approach was implemented to locate a single principal anatomical  
428 region and return its position as a bounding box. Compared with direct segmentation of the entire  
429 fetal head contour, object detection offers a simpler and more computationally efficient method for  
430 isolating the region required by the subsequent Confirm and Measure stages.

431 The Find module utilized the compact YOLO26n object detector, which was trained on full  
432 ultrasound frames and cranial bounding boxes derived from expert fetal-head annotations (37,38).  
433 YOLO26n was selected due to its high reported accuracy in related object detection studies, as well  
434 as its small model size and single-stage detection design. These characteristics align with the  
435 objective of developing a lightweight pipeline capable of localizing the fetal head without  
436 necessitating a large computational model (39,40).

437 During inference, the Find module returned the cranial bounding box with the highest confidence  
438 level, provided that the detection confidence was at least  $\tau_F = 0.25$ . This threshold was  
439 implemented to exclude low-confidence detections while maintaining measurement availability and  
440 was not intended as a clinically validated cutoff. If no detection met the threshold, the frame  
441 remained in the evaluation denominator, the output was recorded as unavailable, and no biometric  
442 measurement was calculated. Thus, the Find module both localized the fetal-head region and  
443 determined whether sufficient localization evidence was present for measurement to proceed. Full  
444 training settings are provided in Additional file 1.

445 **Confirm stage for anatomical-evidence assessment**

446 Confirm was developed to determine whether specific internal cranial structures were visible within  
447 the fetal-head region identified by Find. It evaluated the CSP, a clinically relevant small-anatomy  
448 exemplar, and the LV, as supplementary intracranial evidence. The purpose was to report the  
449 anatomical evidence accompanying each measurement, ensuring that biometric values were  
450 interpreted in the context of the supporting image content.

451 The cranial box generated by Find was expanded to retain surrounding anatomical context and  
452 clipped as needed to remain within the image boundary. The resulting region was converted to  
453 grayscale, denoised, and locally contrast-enhanced, to reduce irrelevant image content and increase  
454 the visibility of the small CSP and LV structures, while maintaining a traceable link to the original  
455 ultrasound frame. The enhanced crop was intended to facilitate model assessment and subsequent  
456 review, rather than to generate anatomical evidence not present in the source image.

457 Confirm was developed using two related training pathways. The primary pathway employed a  
458 staged curriculum to train the CSP/LV evidence model described in the main analysis. The  
459 secondary pathway utilized human-curated structure boxes and a prompt-guided segmentation  
460 model to develop a separate, compact implementation.

461 The prompt-guided pathway was necessary because the available annotations indicated only the  
462 approximate locations of CSP and LV using bounding boxes, without expert-drawn pixel-level  
463 boundaries. Each curated box was provided to Segment Anything Model 2–Small (SAM2-Small),  
464 a general-purpose segmentation model that uses spatial prompts to estimate which pixels within the  
465 specified region likely belong to the structure. SAM2-Small generated provisional CSP and LV  
466 masks at tight, contextual, and wider regional crop scales. These masks offered more detailed  
467 training targets than bounding boxes alone, but remained model-assisted approximations rather than  
468 independent expert annotations. SAM2-Small was used exclusively during training-data  
469 preparation and was not included in ATLAS during test inference.

470 The staged curriculum was implemented to address two challenges common in small-structure  
 471 learning for medical images (Table 2). First, CSP and LV occupy only a small portion of the cranial  
 472 image. Without targeted training, a model may primarily learn from the larger background area,  
 473 resulting in apparently favorable overall performance but inconsistent detection of the structures.  
 474 Second, the available structure labels consisted of approximate boxes and model-generated masks,  
 475 introducing uncertainty regarding their boundaries. The curriculum therefore progressed from broad  
 476 cranial context to structure-specific learning, followed by focused training on challenging  
 477 examples, integration of the CSP and LV specialist models, and final high-resolution refinement.  
 478 This sequence enabled the model to first localize the structures within the fetal head before focusing  
 479 on their smaller and more ambiguous visual features.

480

481 Table 2. Staged learning curriculum used to develop the Confirm anatomical-evidence model

| Phase                               | Training focus                                                                   | Purpose                                                                                                                 |
|-------------------------------------|----------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------|
| S0: Cranial context                 | Full localized fetal-head region                                                 | Learn the broader anatomical context in which CSP and LV are expected to appear.                                        |
| S1a–S1b: Structure specialization   | Separate CSP and LV training streams                                             | Allow the model to learn the distinct appearance and location of each structure without one class dominating the other. |
| S2: Joint evidence learning         | Combined CSP and LV examples across cranial regions                              | Integrate information from both structures into a common evidence model.                                                |
| S3a–S3b: Difficult-example learning | Ambiguous, small, and poorly contrasted CSP and LV examples at higher resolution | Reduce background dominance and improve recognition of less clearly represented structures.                             |
| S4: Specialist integration          | Combine the CSP- and LV-specialist model parameters                              | Retain complementary knowledge learned by the two specialist streams within one model.                                  |
| S5: Final refinement                | High-resolution training with conservative parameter updates                     | Refine structure localization while reducing the risk of disrupting previously learned cranial context.                 |

482 **Note:** The curriculum phases describe the order in which training information was introduced. Detailed image  
 483 resolution, augmentation, learning-rate, and checkpoint settings are provided in Additional file 1.

484 Confirm additionally produced an anatomical evidence parameter (AEP), an exploratory single-  
 485 pass score retained in the output record to provide an exploratory ordering score derived from the  
 486 confirmed-anatomical (e.g., CSP) output. It was computed entirely from fields already produced by  
 487 Confirm and therefore required no additional forward pass, model, or annotation:

488 
$$AEP_i = pCSP_i(1 - H_i)$$

489 The Confirm CSP confidence is  $p_{CSP,i}$ . For CSP-detected frames, ( $H_i = 0.3$ ), and therefore  
490 ( $AEP_i = 0.7 \cdot p_{CSP,i}$ ) for frames without a CSP detection, ( $p_{\{CSP\},i} = 0$ ) and consequently  $AEP_i =$   
491 0. AEP is therefore a monotonic rescaling of CSP confidence within CSP-detected frames rather  
492 than an independent estimate of image quality or uncertainty. Its function was confined to the task  
493 of ordering, sampling, or filtering the output for retrospective analysis and possible review and it  
494 was not employed to classify images as clinically adequate or inadequate.

495 Confirm's primary outputs are categorical, since a detected or not-detected status divides the  
496 outputs into two broad categories but does not allow for ranking within those categories. In a system  
497 designed to enable retrospective audit and prioritized human review, the actual issue is usually not  
498 whether there is evidence but rather which outputs should be given priority. By including a single  
499 scalar value together with the measurement and its source, the outputs can be ordered, sampled or  
500 filtered without having to re-run the inference or refer to the model. As this scalar is based on  
501 existing fields it brings about no additional computational cost.

## 502 **Measure stage for biometric calculation**

503 Measure operationalized the measurement-geometry and physical-scale components of the ATLAS  
504 evidence pathway. It calculated BPD, OFD, and HC from explicitly retained fetal-head geometry  
505 and pixel spacing. This study used two measurement techniques to address design ablation and real-  
506 world data limitations: a segmentation-based method and a direct Find-box measurement method.  
507 The segmentation-based method predicted a head mask from the ultrasound frame and derived  
508 measurements from an ellipse fitted to the mask. The direct Find-box method used the bounding  
509 box dimensions from Find and required no further image analysis. These methods were compared  
510 to determine if a more detailed representation of fetal-head shape improved biometric accuracy  
511 beyond what Find provided. In the ATLAS inference pathway, Measure was executed only when  
512 Find returned an eligible cranial detection. Find served as an availability gate; if it did not identify  
513 a cranial region at the required confidence threshold, no measurement was reported.

#### **Segmentation-based measurement method**

Segmentation builds upon established biometry approaches that typically employ direct segmentation without localization or additional contextual evidence such as CSP or LV (41–44). This method estimates the head boundary, rather than providing only a general location or a directly predicted numerical value. The predicted boundary can be visualized on the ultrasound image for inspection and used to fit an ellipse representing the approximately oval shape of the fetal head. It utilized a lightweight YOLO26n-seg model, which processed the original full ultrasound frame to predict pixels corresponding to the fetal head, thereby generating a fetal-head mask.

The predicted mask may occasionally contained multiple disconnected regions, which may result from image noise or uncertainty at the head boundary, causing the model to identify small areas outside the principal head region. To address this, the largest valid connected region was retained for measurement, based on the expectation that the fetal head would constitute the dominant continuous region in a successful prediction. This is supported by morphology and denoising approaches in segmentation works (4,42). Smaller disconnected regions were considered likely segmentation artifacts and excluded from analysis.

An ellipse was fitted to the boundary of the retained region. The shorter fitted-ellipse axis was assigned to BPD, representing the transverse width of the fetal head, and the longer axis was assigned to OFD, representing its anteroposterior length. Assigning the axes according to their relative lengths ensured consistent identification of BPD and OFD regardless of fetal-head orientation in the image.

The dimensions of the fitted ellipse were initially expressed in pixels. These values were converted to millimeters using the recorded pixel spacing, which specifies the physical distance represented by a single image pixel:

$$BPD (mm) = \text{shorter ellipse axis (pixels)} \cdot \text{pixel spacing} \left( \frac{mm}{pixel} \right)$$

$$OFD (mm) = longer\ ellipse\ axis\ (pixels) \cdot pixel\ spacing\ \left(\frac{mm}{pixel}\right)$$

$$HC = \pi \left[ 3(a + b) - \sqrt{(3a + b) \cdot (a + 3b)} \right]$$

HC was subsequently calculated from the resulting BPD and OFD values using the prespecified ellipse-perimeter equation. After identifying the boundary, the final measurements followed a reproducible sequence: fitting an ellipse to the predicted contour, assigning its shorter and longer axes to BPD and OFD, converting the axes from pixels to millimeters, and calculating HC from these physical dimensions. Each measurement could be traced to the source frame, predicted mask, fitted ellipse, pixel-spacing record, and calculation equation.

#### **Direct find-box measurement method**

A direct Find-box method was evaluated as an alternative implementation of the Measure function. This approach did not process an image or apply an additional segmentation model. Instead, it calculated biometric values directly from the cranial bounding box produced by Find. Similar methods have been utilized in related one-stage detection-based studies that emphasize computational efficiency (32,43,45).

The shorter dimension of the Find box was assigned to BPD, and the longer dimension to OFD. Both dimensions were converted from pixels to millimeters using the recorded pixel spacing, after which HC was calculated using the same ellipse-perimeter equation as in the segmentation-based method. Once Find produced the box, the measurements followed directly from the box dimensions, pixel-spacing record, and prespecified equations.

#### **System output: Structured audit record and dual-context visualization**

For each ultrasound frame, ATLAS generated a structured audit record that linked the biometric result to the underlying information used in its derivation. This design was motivated by the distinction between automated measurement generation and subsequent quality review. Figure 3 presents representative outputs from the evaluation result. The audit record included the source

image and de-identified benchmark identifiers, the pixel-spacing value and its source (either image metadata or the prespecified fallback), the model-checkpoint hash, the Find bounding box and confidence score, the Confirm status for CSP and LV evidence, the head geometry used by Measure, the resulting BPD, OFD, and HC values, and an indicator of measurement availability. This structure enabled each reported value to be traced to the corresponding image, model version, anatomical evidence assessment, measurement geometry, and image-scale information. The structured record retained these fields as separate evidence states rather than combining them into an automated binary judgement.

The accompanying visualization displayed the analysis in two complementary views. The full ultrasound frame maintained the original acquisition context and indicated the location identified by Find for the fetal head. A focused view of the localized cranial region facilitated inspection of smaller internal structures and measurement geometry. Image enhancement was applied solely to improve the visibility of existing information in the source image and was not used to infer the presence of obscured or absent anatomical structures.

## **External evaluation**

External evaluation assessed whether the trained Find and segmentation-based Measure models performed effectively on the unseen FP and UCL datasets without retraining. The Find and segmentation-based Measure models were applied without modification to the patient-exclusive FP and UCL test sets using a zero-shot learning approach (46,47). No test images or subjects from either dataset were used for model development, threshold selection, or adaptation.

Each external dataset provided expert landmarks defining BPD and OFD. These landmarks were used to construct the reference cranial box and calculate the reference BPD and OFD values. Reference HC was then calculated from the landmark-derived BPD and OFD using the same prespecified ellipse-perimeter equation applied to the model outputs. The reference and model-derived HC values were therefore based on the same geometric definition.

587 The external datasets enabled evaluation of the Find and both Measure methods, but not Confirm.  
588 Find was assessed based on detection availability and the overlap between the predicted cranial box  
589 and the landmark-derived reference box. The direct Find-box Measure calculated BPD, OFD, and  
590 HC from the predicted Find-box dimensions and pixel spacing. Segmentation-based Measure  
591 analyzed the original full ultrasound frame, predicted a fetal-head contour, fitted an ellipse to the  
592 contour, and calculated the biometric parameters from the fitted axes and pixel spacing. External  
593 evaluation thus assessed whether each implementation produced accurate biometric values under  
594 the same Find-gated availability criteria.

595 For zero-shot external evaluation of segmentation-based Measure, the model was applied to each  
596 FP and UCL test frame. The largest valid predicted contour was retained and fitted with an ellipse.  
597 The shorter and longer ellipse axes were converted to BPD and OFD using frame-level pixel  
598 spacing, and HC was calculated using the same ellipse equation used internally. External references  
599 were derived from expert BPD and OFD landmark endpoints and the corresponding ellipse-derived  
600 HC. Dice and mask IoU were not calculated because the external sources did not provide compatible  
601 expert fetal-head masks. No pseudo-masks were generated for external evaluation. This evaluation  
602 determined whether each measurement method produced accurate biometric values on the external  
603 images.

604 For the direct Find-box method evaluation, two baselines were included to assess whether the  
605 performance of the method could be attributed to average fetal size or image position. The median-  
606 measurement baseline assigned the same median reference BPD and OFD values to every test  
607 frame, representing an estimate based solely on typical fetal-head size. The centred median-box  
608 baseline placed a box with the median reference width and height at the center of each test image,  
609 representing an estimate based on typical fetal-head size and the assumption of central head  
610 location. The models and baselines were evaluated on the same, unchanged FP and UCL test sets.

611 These comparisons determined whether the learned Find model provided frame-specific  
612 localization and measurement information beyond the two baseline assumptions.

### 613 **Adaptation and aggregation**

614 Robust medical imaging AI typically requires training data from multiple institutions. Models  
615 developed using data from a single institution may perform differently in other settings due to  
616 variations in patient populations, ultrasound systems, acquisition practices, image characteristics,  
617 and annotation procedures. Access to diverse data can improve model development and provide  
618 evidence for cross-site generalizability. However, pooling clinical images in a central environment  
619 is often challenging because health data are governed by separate institutional, ethical, legal, and  
620 technical arrangements (48,49). These constraints present a significant barrier to the development  
621 of multi-institutional medical imaging AI.

622 Federated learning has been proposed as a response to this barrier (50,51). In a conventional  
623 federated workflow, participating institutions retain local data and exchange model updates instead  
624 of transferring patient images. This approach enables collaborative learning across decentralized  
625 datasets. Against this background, this research also examined the feasibility of ATLAS to be  
626 transferred and deployed on specific sites with different cohort data, and aggregated in federated  
627 learning approach for privacy-preserving collaborative learning.

628 This evaluation was deliberately restricted to the Find stage due to considerations of leverage and  
629 reference-standard availability. The Find stage determines measurement availability and defines the  
630 region from which all downstream evidence and geometry are derived. Consequently, any  
631 degradation introduced by adaptation at this stage propagates to all subsequent stages, making it the  
632 component whose adaptation has the greatest downstream impact. The restriction is further  
633 necessitated by the external data: FP and UCL provide expert landmarks but lack CSP or LV  
634 evidence labels, preventing adaptation or evaluation of the Confirm stage on these cohorts.  
635 Additionally, their head references are derived from landmarks rather than expert pixel-level masks.

636 As a result, fine-tuning a segmentation model for segmentation-based measurement would optimize  
637 the model toward a derived target and confound the effect of adaptation with the geometry of that  
638 derivation. Therefore, adaptation and aggregation tests are restricted to the Find stage to isolate the  
639 mechanism under investigation, specifically, whether site-specific parameter changes preserve the  
640 clinical measurement, without introducing a second, more weakly supervised stage into the  
641 comparison. Parameter aggregation is also most tractable for the compact detector, where task  
642 vectors are structurally aligned across specialists.

### 643 **Site-specific fine-tuning**

644 The site-specific evaluation investigates whether the trained Find model could be fine-tuned  
645 separately using FP and UCL development data and whether the resulting site-specific adaptations  
646 could subsequently be combined without accessing the source images during aggregation. The  
647 analysis tested the feasibility and measurement consequences of combining knowledge learned  
648 separately from external datasets while avoiding centralized image pooling at the aggregation stage.

649 All models in this analysis were initialized from the same trained model. Fine-tuning was employed  
650 to develop site-specific models, where an already trained model is further trained using data from a  
651 new dataset, enabling adaptation to the characteristics of that dataset without relearning the task  
652 from the beginning (50). This common starting point facilitated comparison and subsequent  
653 combination of the changes learned from FP and UCL.

654 One copy of this model was fine-tuned on the patient-exclusive FP development partition to produce  
655 the FP specialist, while a separate copy was fine-tuned on the patient-exclusive UCL development  
656 partition to produce the UCL specialist. A third copy was fine-tuned on the combined FP and UCL  
657 development partitions to produce the pooled-external model. The pooled model represented the  
658 conventional scenario where both development datasets were available within a single training  
659 environment, whereas the two specialists represented site-separated model adaptation. No external  
660 test subjects contributed to fine-tuning, checkpoint selection, or model aggregation.

Each adapted model was evaluated on the unchanged FP and UCL test sets. Performance was assessed using both the overlap with the reference cranial box and the accuracy of BPD, OFD, and HC calculated from the predicted box using the direct Find-box method. Evaluating both outcomes was necessary because increased box overlap does not guarantee that the box width and height, and thus the resulting biometric measurements, remain accurate. The analysis therefore tested whether site-specific fine-tuning improved localization while preserving downstream clinical measurements.

#### **Aggregation of the site-specific models**

TIES-Merging was subsequently used to combine the FP and UCL specialists into a single Find model. Combining locally adapted models offers a potential approach to integrating knowledge learned across institutions without transferring source images during aggregation. However, previous studies indicate that differences between institutional datasets can cause locally learned model updates to conflict, potentially reducing performance when updates are combined by simple parameter averaging (52,53). This presents challenges for multi-institutional medical imaging AI.

TIES-Merging was selected because it was specifically developed to address interference between models derived from a common starting checkpoint (54). The method is based on evidence that directly averaging separately adapted models can discard useful information when parameter changes include many small, potentially unimportant updates or changes in opposing directions. TIES-Merging addresses these issues by retaining larger parameter changes, resolving directional disagreements, and combining only changes consistent with the selected direction. The method has demonstrated advantages over conventional merging approaches across various tasks, domains, model architectures, and fine-tuning settings, although it was not originally developed for fetal ultrasound or privacy-preserving medical AI (55).

TIES-Merging was applied because the differences between the original model parameters and the resulting FP and UCL specialist parameters represented the adaptations learned from each external

dataset. These parameter changes were combined in three steps. First, Trim retained the largest 20% of changes and removed smaller changes that were less likely to represent meaningful adaptation. Second, Elect Sign identified the prevailing direction of change for each parameter and removed site-specific changes that conflicted with that direction. Third, Disjoint Merge combined the remaining non-conflicting changes. The combined changes were then applied to the fixed HC18 checkpoint to construct a single merged Find model.

The reconstructed model was reloaded and evaluated on the original FP and UCL test sets. Performance was assessed using cranial-box overlap (Find metrics) as well as BPD, OFD, and HC, which were calculated directly from the predicted Find box (Measure metrics). This measurement-aware evaluation was conducted because successful parameter aggregation or improved localization overlap does not necessarily ensure preservation of the box dimensions required for accurate fetal-head biometry. This experiment serves as a preliminary assessment of image-free model aggregation, motivated by challenges associated with centralized multi-institutional training.

### **Clinical expert quality review**

Quantitative evaluation was complemented by clinical review to assess whether the stage-specific outputs of ATLAS were clinically interpretable and corresponded to clinically meaningful features of fetal-head biometry. A Maternal Fetal Medicine (MFM) fellow who was not part of the author group reviewed ATLAS test-result images and compared the system outputs with clinician-labelled assessments of the corresponding images based on ISUOG-relevant criteria for fetal-head biometric imaging. The review considered cranial-plane suitability, visibility of relevant intracranial anatomy, measurement geometry. Discordant or technically questionable cases were examined to identify whether apparent computational failure or uncertainty corresponded to clinically recognisable limitations in image or measurement quality.

A senior MFM specialist and medical director (Author 3) subsequently reviewed representative ATLAS outputs, selected cases, and the interpretation of the observed patterns. This senior review

711 was used to assess the clinical relevance and potential usefulness of the evidence-traceability  
712 approach and to contextualize the findings within fetal-ultrasound practice. The senior MFM  
713 reviewer did not independently review the complete test dataset, and this component was therefore  
714 treated as clinical interpretation rather than an independent performance assessment. The clinical  
715 review was intended to assess validity, interpretability, and the clinical relevance of the evidence  
716 states produced by ATLAS.

#### 717 **Evaluation metrics**

718 The evaluation was structured to align with the specific objectives of each ATLAS stage: Find,  
719 Confirm, and Measure. Table 3 presents the corresponding evaluation metrics.

720 Detection availability assessed whether the Find stage identified an eligible cranial region suitable  
721 for subsequent analysis. Intersection over Union (IoU) quantified the overlap between predicted  
722 and reference cranial boxes. Since a box could exhibit high overlap despite inaccurate width or  
723 height, IoU was interpreted in conjunction with biometric outcomes to ensure consistency and  
724 integrity.

725 The performance of the Confirm stage was evaluated using precision, recall, F1 score, and accuracy.  
726 Precision measured the frequency with which reported CSP or LV evidence matched the  
727 corresponding positive label, while recall quantified the detection rate of labelled evidence. The F1  
728 score summarized the balance between precision and recall, and accuracy reflected overall  
729 agreement. These metrics were calculated based on box-derived CSP/LV evidence from the data  
730 integrity layer, representing agreement with the study's evidence labels rather than accuracy relative  
731 to independent expert assessment of anatomical visibility or standard-plane suitability.

732 The measure was evaluated using both error magnitude and error direction. BPD and OFD accuracy  
733 were summarised as MAE in millimeters. HC was evaluated using absolute percentage error to  
734 account for differences in fetal-head size. Mean signed error was also reported to identify systematic

735 bias, with negative values indicating average under-measurement and positive values indicating  
736 average over-measurement.

737 **Joint measurement criterion**

738 Average errors do not indicate the frequency with which individual outputs satisfy multiple  
739 measurement requirements. Therefore, a joint-tolerance criterion was prespecified. An output  
740 satisfied this criterion if the BPD absolute error was no greater than 2 mm and the HC absolute  
741 percentage error was no greater than 5%:

742 
$$|\widehat{BPD} - BPD_{reference}| \leq 2 \text{ mm}$$

743 and

744 
$$\frac{|\widehat{HC} - HC_{reference}|}{HC_{reference}} \cdot 100 \leq 5\%$$

745 The joint-tolerance rate represented the proportion of reportable outputs that satisfied both limits.  
746 Requiring both conditions ensured that superior performance in one biometric parameter did not  
747 mask substantial error in the other. The limits were chosen as clinically interpretable bounds,  
748 informed by fetal-head measurement standardization and reproducibility literature; they do not  
749 represent diagnostic, treatment, or device-approval thresholds (1).

750 **Confirm: Anatomical-evidence-stratified analysis**

751 A secondary analysis assessed whether measurement performance varied according to the CSP  
752 evidence reported by Confirm. Outputs were categorized into CSP-detected and CSP-not-detected  
753 groups, and BPD error, HC error, and the joint-tolerance rate were summarized for each group.  
754 Patient-cluster bootstrap resampling was again employed to ensure that repeated frames from the  
755 same patient remained grouped. This analysis aimed to determine whether the Confirm evidence  
756 state provided additional context for interpreting measurement performance.

## 757 **Measure: Comparison of the two measure methods**

758 The segmentation-based and direct Find-box measure methods were evaluated using the same  
759 HC18 internal test set. A paired design was employed, as both implementations generated  
760 measurements for identical images and were compared with the same reference values. This  
761 approach enabled more direct attribution of performance differences to the methods used for  
762 estimating head geometry.

763 For each frame, paired differences between the two implementations were calculated for BPD  
764 absolute error, OFD absolute error, HC absolute percentage error, signed measurement error, and  
765 joint-tolerance status. Mean paired differences and their 95% confidence intervals (CI) were  
766 estimated using 10,000 patient-cluster bootstrap resamples. Patients served as the resampling unit,  
767 ensuring that all frames from the same patient remained grouped and accounting for dependence  
768 between repeated frames. This procedure was applied to all paired comparisons between the  
769 segmentation-based and direct Find-box measure methods.

## 770 **Computational footprint**

771 Runtime was measured on a single workstation, and the checkpoint footprint was calculated using  
772 the exact model files used in the evaluated pathway. These outcomes describe computational  
773 performance under the tested conditions and do not establish feasibility on an ultrasound console  
774 or edge device. The exploratory CSP-confidence analysis is reported in Additional file 1 and was  
775 not considered an independently validated uncertainty method.

776

## 777 **Results**

778 The results are structured to reflect the progression from data preparation through external and  
779 adaptation-aggregation evaluation. The integrity and traceability of the data are reported first,  
780 followed by the internal performance of individual ATLAS components. Ablation results for each  
781 ATLAS component are then presented, followed by an assessment of transfer to external datasets.

782 Subsequent analyses evaluate site-specific model adaptation and aggregation, as well as the  
783 computational implementation profile. Finally, ATLAS is compared with representative fetal-  
784 ultrasound AI approaches.

## 785 **Data Integrity Layer Performance**

786 The data-integrity layer generated 3,790 frames from 1,840 patients across the prespecified training,  
787 validation, and test partitions. A programmatic audit was conducted after dataset construction, with  
788 results provided in Additional File 1. Patient exclusivity was verified by comparing patient  
789 identifiers across each pair of partitions; no patient appeared in more than one partition. The frame-  
790 level ledger contained 3,790 unique identifiers, with no duplicated source records. On-disk checks  
791 confirmed a one-to-one correspondence between images and label files in each partition: 2,629  
792 images and 2,629 labels in training, 575 and 575 in validation, and 586 and 586 in testing. Empty  
793 structure-label files were retained as explicit negative examples rather than classified as missing  
794 data. Observed patient and frame counts matched the prespecified partition denominators exactly.  
795 The locked internal test cohort included all 586 frames from 279 patients, with no post hoc  
796 exclusions or subsampling.

797 A controlled fault-injection experiment demonstrated the role of data integrity layer as  
798 complementary validity controls (a brief description of the ablation is provided in Table 3, with the  
799 full detail in Additional File). The complete integrity layer detected all five deliberately introduced  
800 defects: patient overlap across partitions, a duplicated frame identifier, a missing label, invalid pixel  
801 spacing, and disagreement between observed and prespecified cohort counts. When an individual  
802 integrity check was disabled, the specific defect targeted by that check was not detected, while the  
803 remaining four defects continued to be identified. This experiment does not indicate that the  
804 integrity layer improved model accuracy. Instead, it demonstrates that each control addressed a  
805 distinct failure mode and that no single check could substitute for the complete set.

806 Table 3. Ablation result of data integrity layer

| Integrity-control configuration      | Cross-partition patient overlap | Duplicate frame identifier | Missing label | Invalid pixel spacing | Cohort-count mismatch | Defects detected among those assessed |
|--------------------------------------|---------------------------------|----------------------------|---------------|-----------------------|-----------------------|---------------------------------------|
| Complete data-integrity layer        | Detected                        | Detected                   | Detected      | Detected              | Detected              | 5/5                                   |
| Patient-exclusivity check removed    | Not assessed                    | Detected                   | Detected      | Detected              | Detected              | 4/4                                   |
| Duplicate-identifier check removed   | Detected                        | Not assessed               | Detected      | Detected              | Detected              | 4/4                                   |
| Label-presence check removed         | Detected                        | Detected                   | Not assessed  | Detected              | Detected              | 4/4                                   |
| Pixel-spacing validity check removed | Detected                        | Detected                   | Detected      | Not assessed          | Detected              | 4/4                                   |
| Cohort-count reconciliation removed  | Detected                        | Detected                   | Detected      | Detected              | Not assessed          | 4/4                                   |

807 **Note:** A blank result in the experimental output has been presented as “Not assessed,” since the corresponding control  
808 was intentionally disabled. This should not be taken to mean that the check has passed or that the defect is absent.  
809 “Cohort-count mismatch” indicates disagreement between the observed numbers of patients or frames and the  
810 prespecified evaluation counts.

## 811 Internal Component Performance

812 Figure 3 shows the ATLAS result on internal test on intended pathway: CSP evidence detected,  
813 accurate geometry, and both tolerance limits met.

## 814 Find Component

815 The Find component localized the fetal cranial field in all 586 test frames, achieving 100%  
816 identification of measurement availability. The mean and median intersection over union (IoU)  
817 were 0.962 and 0.968, respectively (Table 4). These results indicate that the system consistently  
818 identified a cranial region closely matching the human expert-derived reference box.

## 819 Confirm Component

820 The Confirm component detected CSP evidence in 295 of 586 frames (50.3%) and LV evidence in  
821 208 of 586 frames (35.5%) (Table 4). Compared to the box-derived evidence ledger from, CSP  
822 precision was 0.888, recall 0.900, F1 score 0.894, and accuracy 0.894. LV precision was 0.846,  
823 recall 0.793, F1 score 0.819, and accuracy 0.867. These results demonstrate that Confirm  
824 reproduced the curated CSP and LV evidence representation on unseen patients. However, these

findings do not establish accuracy against independent expert assessment of standard-plane suitability.

Table 4. Internal component performance

| Stage       | Endpoint                                 | Result                              | Interpretation                                                                                  |
|-------------|------------------------------------------|-------------------------------------|-------------------------------------------------------------------------------------------------|
| Find        | Availability;<br>mean IoU;<br>median IoU | 586/586 (100%);<br>0.962;<br>0.968  | All locked-test frames yielded a well-overlapping cranial field.                                |
| Confirm–CSP | Precision;<br>recall;<br>F1;<br>accuracy | 0.888;<br>0.900;<br>0.894;<br>0.894 | The model reproduced the curated CSP evidence representation on unseen patients.                |
| Confirm–LV  | Precision;<br>recall;<br>F1;<br>accuracy | 0.846;<br>0.793;<br>0.819;<br>0.867 | The model reproduced the curated LV evidence representation, with lower agreement than for CSP. |

Note: Higher values are favorable. Confirm metrics quantify consistency with a box-derived pseudo-reference ledger, not independent expert standard-plane accuracy.

Consistent with these findings, CSP evidence was not detected in any of the 9 test frames with a reference head circumference below 110 mm (Additional File 1). Detection rates were 18.2% and 28.0% in the 110–129 mm and 130–154 mm bands, respectively, compared with 52.9% in frames above 155 mm. These early-gestation strata are small (45 frames, 7.7% of the test cohort) and have wide, overlapping confidence intervals. No association was evident across the cohort as a whole. Therefore, this observation is consistent with, but does not establish, the developmental emergence of the CSP as a contributor to non-detection.

## Measure Component

Both Measure implementations produced low fetal-head measurement errors. However, the direct Find-box geometry method demonstrated stronger overall biometric performance on the internal test set (Table 5). The primary advantage of the direct pathway was greater accuracy for BPD and a higher proportion of outputs meeting both the prespecified BPD and HC accuracy limits.

The segmentation-based Measure method produced a valid fitted ellipse for all 586 test frames. The MAE was 1.069 mm for BPD, 1.510 mm for OFD, and 1.242% for HC. Overall, 493 of 586 outputs (84.1%) were within both accuracy limits, defined as a BPD error of no more than 2 mm and an HC

error of no more than 5%. The mean signed error indicated that this pathway tended to underestimate BPD by 0.840 mm and HC by 0.608 mm.

The direct Find-box geometry method also produced measurements for all 586 frames and outperformed the segmentation-based approach on the combined biometric endpoint. Its MAE was 0.779 mm for BPD, 1.484 mm for OFD, and 1.259% for HC. A total of 537 of 586 outputs (91.6%) met both accuracy limits. In paired analysis, direct geometry reduced BPD error by 0.291 mm compared with segmentation (95% CI 0.225–0.355) and increased the proportion meeting both limits by 7.5% (95% CI 5.2–9.9). Differences in OFD and HC error were small and statistically inconclusive. Direct geometry also demonstrated substantially less systematic BPD underestimation, with a mean signed error of –0.154 mm.

These findings indicate that the benefit of direct geometry was concentrated in BPD and in the combined BPD–HC accuracy criterion, rather than across all measurements. Within the same Find–Confirm framework, the simpler Find-box implementation provided more reliable numerical Measure outputs for the available labels and evaluation endpoints. The segmentation pathway retained practical value by producing an explicit head contour and fitted ellipse for display and review.

Table 5. Internal comparison of the two Measure methods on the internal test set

| Measure implementation                | BPD MAE  | OFD MAE  | HC MAE | BPD signed bias | Outputs meeting both accuracy limits |
|---------------------------------------|----------|----------|--------|-----------------|--------------------------------------|
| Segmentation-based measurement method | 1.069 mm | 1.510 mm | 1.242% | –0.840 mm       | 493/586 (84.1%)                      |
| Direct Find-box measurement method    | 0.779 mm | 1.484 mm | 1.259% | –0.154 mm       | 537/586 (91.6%)                      |

**Note:** Lower MAE indicates better performance. The combined accuracy criterion represents the proportion of outputs with BPD error  $\leq 2$  mm and HC error  $\leq 5\%$ . Confidence intervals were estimated using 10,000 patient-cluster bootstrap resamples.

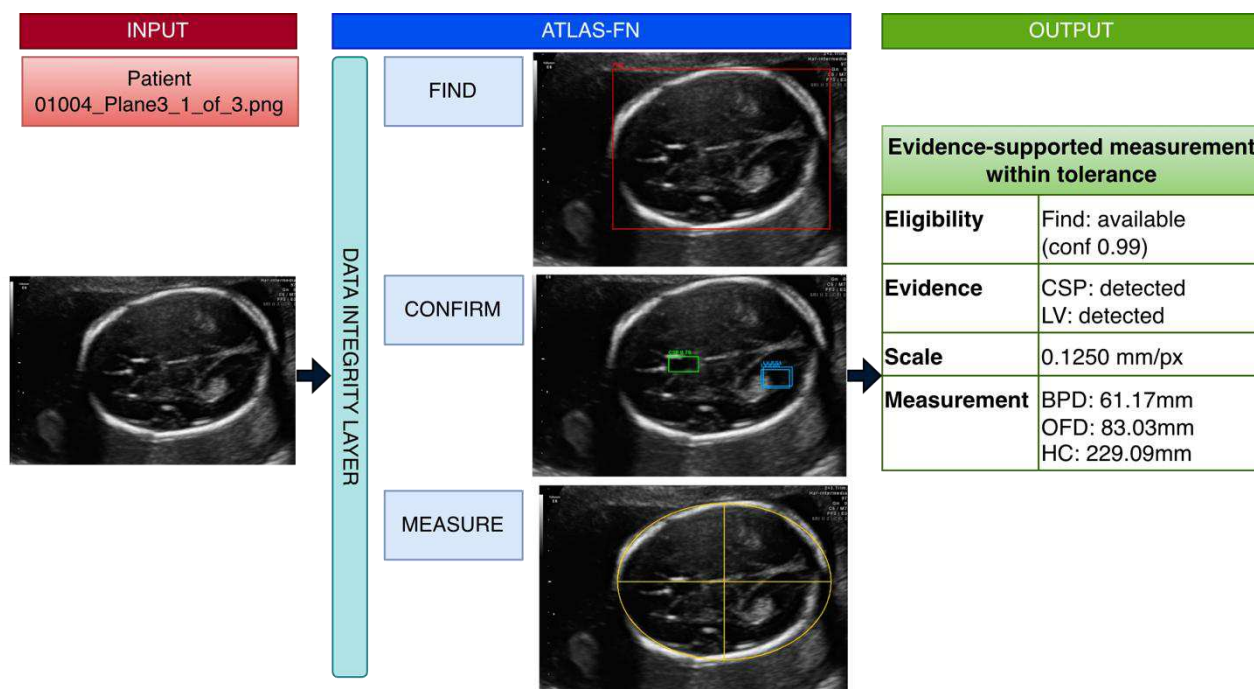


Figure 3. Structured output records for representative test frame

### Ablation Analysis Results

The ablation analysis demonstrated that Find, Confirm, and Measure each contributed distinct functions to the ATLAS pathway, rather than independently enhancing the same accuracy metric. Specifically, Find determined the availability of eligible cranial localization, Confirm documented selected anatomical evidence, and Measure converted image geometry into BPD, OFD, and HC. The impact of removing or bypassing these functions was therefore evaluated using measurement availability, biometric error, anatomical-evidence availability, and the proportion of outputs meeting both prespecified accuracy thresholds.

As shown in Table 6, removing Confirm did not alter the numerical measurements produced by either Measure implementation. For segmentation-based Measure, the BPD mean absolute error remained 1.069 mm, HC error remained 1.242%, and 493 of 586 outputs (84.1%) met both accuracy thresholds. For direct Find-box Measure, BPD error remained 0.779 mm, HC error remained 1.259%, and 537 of 586 outputs (91.6%) met both thresholds. However, removing Confirm eliminated the CSP and LV evidence fields from all outputs. This outcome was anticipated, as Confirm does not modify measurement geometry in the current architecture. Its primary role is to preserve anatomical context and reviewability, rather than to numerically adjust the measurement.

883 Bypassing the Find availability gate for full-frame segmentation did not affect internal measurement  
884 performance. Find identified an eligible cranial localization for all 586 HC18 test frames, resulting  
885 in no Find-unavailable group to assess changes in measurement availability. This finding does not  
886 indicate that Find is unnecessary. Instead, the internal cohort was unable to evaluate the protective  
887 function of the Find unavailable-output state. A comprehensive Find ablation would require images  
888 where localization fails or includes distractor anatomy, as well as a separately defined replacement  
889 for the Find-dependent Confirm and direct-measurement pathways.

890 Eliminating Measure resulted in the complete absence of biometric output. While Find and Confirm  
891 continued to provide cranial localization and anatomical-evidence information, no BPD, OFD, or  
892 HC values were generated. In contrast, altering the source of measurement geometry produced a  
893 quantifiable effect. Direct Find-box geometry reduced BPD error by 0.291 mm compared to  
894 segmentation-based geometry (95% CI 0.226–0.357) and increased the proportion of outputs  
895 meeting both accuracy thresholds by 7.5% points (95% CI 5.2–10.0). HC error differed by only  
896 –0.017 percentage points, with a confidence interval that included zero. Thus, the primary  
897 numerical impact of the component comparison was concentrated in BPD and the combined BPD–  
898 HC endpoint.

899 Table 6. Ablation of ATLAS components on the internal test set

| Configuration                                            | CSP/LV<br>evidence<br>record<br>retained | Biometry<br>available | BPD<br>MAE  | HC<br>error | Outputs<br>meeting<br>both<br>accuracy<br>limits | Functional interpretation                                                                                                                                                                                   |
|----------------------------------------------------------|------------------------------------------|-----------------------|-------------|-------------|--------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| - Find                                                   | Yes                                      | 586/586<br>(100%)     | 1.069<br>mm | 1.242%      | 493/586<br>(84.1%)                               | No numerical effect because Find was available for every internal test frame. This was not a complete removal of Find. Measurements were unchanged, but CSP/LV evidence was no longer available for review. |
| - Confirm;<br>segmentation-<br>based Measure<br>retained | No                                       | 586/586<br>(100%)     | 1.069<br>mm | 1.242%      | 493/586<br>(84.1%)                               |                                                                                                                                                                                                             |
| - Confirm; direct<br>Find-box<br>Measure retained        | No                                       | 586/586<br>(100%)     | 0.779<br>mm | 1.259%      | 537/586<br>(91.6%)                               | Measurements were unchanged, but CSP/LV evidence was no longer available for review.                                                                                                                        |
| - Measure                                                | Yes                                      | 0/586<br>(0%)         | —           | —           | 0/586<br>(0%)                                    | Localization and anatomical evidence remained available,                                                                                                                                                    |

|                                          |     |                |          |        |                 |                                                                            |
|------------------------------------------|-----|----------------|----------|--------|-----------------|----------------------------------------------------------------------------|
|                                          |     |                |          |        |                 | but no biometric value was emitted.                                        |
| Full FCM with segmentation-based Measure | Yes | 586/586 (100%) | 1.069 mm | 1.242% | 493/586 (84.1%) | Complete pathway with visible head contour and anatomical-evidence record. |
| Full FCM with direct Find-box Measure    | Yes | 586/586 (100%) | 0.779 mm | 1.259% | 537/586 (91.6%) | Complete pathway using Find geometry for biometry.                         |

**Note:** “CSP/LV evidence record retained” indicates that Confirm outputs were available for each frame; it does not mean that CSP or LV was positively detected in every frame. BPD MAE is reported in millimeters, and HC error is the mean absolute percentage error. “Outputs meeting both accuracy limits” indicates BPD error  $\leq 2$  mm and HC error  $\leq 5\%$ . The Find-gate configuration bypassed only the availability gate applied to full-frame segmentation; it did not remove Find from the generation of the existing Confirm outputs.

### **Confirm evidence as an indicator of measurement reliability**

This section examined whether the anatomical evidence reported by Confirm carries information about the reliability of the resulting measurement, and whether that information is sufficient to act upon. Two related questions were addressed in sequence. The first is evidentiary: whether biometric accuracy differed between frames in which CSP evidence was and was not detected. The second is operational: whether any such difference was large enough to justify using Confirm as a reporting gate that withholds measurements when evidence is absent. Both analyses were computed on the same partition of the internal test set, the 295 frames in which Confirm reported CSP evidence and the 291 in which it did not, and therefore represent two views of one comparison: the first expressed as a difference in accuracy between groups, the second as the consequence for output availability across the full cohort.

This section evaluated the interaction of components by comparing biometric accuracy between frames (Measure) with and without CSP evidence (Confirm) as shown in Table 7.

For segmentation-based measurement method, biometric performance was numerically superior when CSP evidence was detected. BPD MAE was 1.037 mm in the CSP-detected group and 1.103 mm in the CSP-not-detected group. HC error was 1.158% and 1.328%, respectively, while the proportion of outputs meeting both accuracy limits was 85.8% compared with 82.5%. However, the patient-cluster CI for these differences included zero, indicating that the observed differences were not statistically significant.

For direct Find-box measurement method, biometric performance was also improved when CSP evidence was detected. BPD MAE increased from 0.703 mm in the CSP-detected group to 0.855 mm in the CSP-not-detected group, a difference of 0.152 mm (95% CI 0.018–0.292). HC error followed a similar trend, increasing from 1.169% to 1.350%, although the confidence interval for this difference included zero. The proportion of outputs meeting both accuracy limits decreased from 93.2% to 90.0%.

Overall, CSP-detected frames demonstrated more favorable measurement results across both Measure implementations. Only the BPD difference for direct Find-box method had a CI excluding zero, indicating that the higher BPD error observed when CSP evidence was not detected was unlikely to be attributable to sampling variation alone. This finding suggests that CSP status was more clearly associated with BPD accuracy in the direct Find-box method, and that Confirm provided an additional evidence-based grouping of the measurement outputs.

Table 7. Internal biometric performance according to CSP evidence detected by Confirm

| Measure implementation     | CSP evidence reported by Confirm | Test frame      | BPD MAE  | HC MAE | Outputs meeting both accuracy limits |
|----------------------------|----------------------------------|-----------------|----------|--------|--------------------------------------|
| Segmentation-based Measure | Detected                         | 295/586 (50.3%) | 1.037 mm | 1.158% | 85.8%                                |
| Segmentation-based Measure | Not detected                     | 291/586 (49.7%) | 1.103 mm | 1.328% | 82.5%                                |
| Direct Find-box geometry   | Detected                         | 295/586 (50.3%) | 0.703 mm | 1.169% | 93.2%                                |
| Direct Find-box geometry   | Not detected                     | 291/586 (49.7%) | 0.855 mm | 1.350% | 90.0%                                |

**Note:** “Not detected” means that Confirm did not identify CSP evidence in the frame, not that the frame or measurement was missing. Lower MAE is favorable. “Outputs meeting both accuracy limits” indicates BPD error  $\leq 2$  mm and HC error  $\leq 5\%$ . CSP status was generated by Confirm and was not independently adjudicated by clinical experts. Bold values indicate the better result within each Measure implementation.

Given that CSP-detected frames produced more accurate measurements, especially for direct geometry, the subsequent consideration is whether Confirm should be operationally implemented to withhold measurements in the absence of supporting evidence. Confirm was thus assessed as an exploratory reporting gate, with performance recalculated both for retained outputs and across the entire test cohort. The use of the full cohort as a denominator is critical for interpretation: conditional accuracy among retained frames will inherently improve when the least certain frames

947 are excluded, making this metric meaningful only when presented alongside the proportion of the  
948 cohort for which measurements remain available.

949 Reframing Confirm from a contextual evidence component to an exploratory reporting gate yielded  
950 only modest improvements in measurement accuracy among retained outputs, while output  
951 availability decreased substantially (see Additional File 1 for detailed results). A CSP-only gate  
952 retained 295 of 586 frames (50.3%). Among these retained frames, segmentation-based BPD error  
953 decreased from 1.069 mm to 1.037 mm, and the proportion meeting both accuracy thresholds  
954 increased from 84.1% to 85.8%. Direct-geometry BPD error decreased from 0.779 mm to 0.703  
955 mm, with the corresponding proportion rising from 91.6% to 93.2%.

956 These improvements in measurement rates must be evaluated in the context of overall coverage.  
957 When all 586 test frames were used as the denominator, accurate segmentation-based outputs were  
958 available for 43.2% of the cohort, and accurate direct Find-box outputs for 46.9%. Therefore, the  
959 observed improvement among retained outputs was partially due to withholding measurements  
960 from nearly half of the test frames.

961 Using either CSP or LV evidence as the reporting criterion increased the number of retained frames  
962 to 348 of 586 (59.4%). Measurement performance among these retained outputs was comparable  
963 to that observed with the CSP-only gate. Segmentation-based BPD error was 1.032 mm, with 85.9%  
964 of retained outputs meeting both accuracy thresholds. Direct-geometry BPD error was 0.704 mm,  
965 with 93.4% meeting both thresholds. However, when all test frames were considered, the  
966 corresponding successful-output rates were 51.0% and 55.5%, respectively.

967 These results suggest that Confirm can be used to identify a smaller subset with marginally  
968 improved average measurement performance; however, this improvement is accompanied by a  
969 substantial reduction in output availability. This analysis demonstrates an accuracy–availability  
970 trade-off, rather than establishing a clinically validated rule for accepting or withholding fetal-head

971 measurements. The final ATLAS design presented in this study does not use Confirm as a reporting  
972 gate, but rather as a tool to enable more informed evaluation.

### 973 **External Model Transfer**

974 External evaluation determined whether the trained Find model added with the two Measure  
975 methods could localize the fetal head and support biometric measurement on unseen FP and UCL  
976 images without retraining.

### 977 **Transferability of the Find model**

978 The Find model localized the fetal cranial region in 879 of 880 FP frames and all 55 UCL frames.  
979 The mean overlap between predicted and reference cranial boxes was 0.913 on FP and 0.884 on  
980 UCL. These findings demonstrate that the source-trained detector transferred effectively to both  
981 external datasets, although localization agreement was lower on UCL.

### 982 **Transferability of the Two Measure Methods**

983 The performance of the two Measure methods varied by dataset. On FP, direct Find-box method  
984 resulted in lower OFD and HC errors and a 3.2% point higher combined-limit rate compared to the  
985 segmentation-based Measure method. The paired BPD difference was minimal and inconclusive.  
986 On UCL, segmentation reduced BPD error by 0.435 mm and HC error by 0.381% points, and  
987 increased the combined-limit rate by 16.7% points, despite producing one fewer measurement.  
988 These results indicate that neither measurement pathway was consistently superior across external  
989 domains.

990 Measurements calculated directly from the predicted Find boxes remained accurate on FP but were  
991 less consistent on UCL. On FP, BPD MAE was 0.927 mm, OFD error was 1.792 mm, and HC error  
992 was 1.244%; 91.2% of outputs met both prespecified BPD and HC accuracy limits. This supports  
993 transfer of Find-derived cranial geometry to an external test cohort of 460 subjects. Performance  
994 was weaker on UCL. BPD MAE was 1.936 mm, OFD error was 3.716 mm, and HC error was  
995 2.843%; 61.8% of outputs met both accuracy limits. UCL contained only 15 test subjects and 55  
996 frames, and 20 frames required fallback pixel spacing. The UCL estimates are therefore less precise

and indicate sensitivity to dataset and image-scale differences rather than definitive failure of external transfer. Across all 934 external frames for which measurements were produced, weighted BPD MAE was 0.987 mm, HC error was 1.338%, and 89.5% met both accuracy limits.

The segmentation-based Measure model was evaluated without retraining on the patient-exclusive FP and UCL test cohorts. External references were established using expert BPD and OFD landmarks and the corresponding ellipse-derived HC. Due to the unavailability of compatible expert fetal-head masks, this analysis focused on biometric agreement rather than external contour accuracy. On FP, the segmentation-based Measure model produced a valid fitted ellipse in 879 of 880 frames. The BPD mean absolute error (MAE) was 0.953 mm, OFD MAE was 1.978 mm, and HC absolute percentage error was 1.508%. A total of 774 out of 879 outputs (88.1%) met both prespecified accuracy limits. On UCL, measurements were produced for 54 of 55 frames, with a BPD MAE of 1.432 mm, OFD MAE of 3.467 mm, HC error of 2.438%, and 43 of 54 outputs (79.6%) meeting both limits.

**Comparison with simple non-learning baselines**  
Performance of non-learning baseline of FP and UCL is calculated for comparison between using the Find-Confirm-Measure pathway of ATLAS or directly measure using the FP- and UCL-based median and centred median box measurements (Table 8). Both baselines performed substantially worse than ATLAS. Assigning the same median fetal-head measurements to every image resulted in BPD errors exceeding 22 mm on FP and 26 mm on UCL. Placing a typically sized box at the center of each image also produced significantly larger localization and measurement errors. These comparisons demonstrate that the external results reflected frame-specific cranial localization rather than reliance on average fetal size or image position.

Table 8. External performance of ATLAS

| Domain | Method                       | Detection | Mean IoU | BPD MAE   | HC MAE  | Outputs meeting both accuracy limits |
|--------|------------------------------|-----------|----------|-----------|---------|--------------------------------------|
| FP     | Baseline: Median measurement | N/A       | N/A      | 22.922 mm | 27.138% | 13.4%                                |
| FP     | Baseline: Centred median box | 100%      | 0.722    | 8.616 mm  | 10.517% | 12.0%                                |

|            |                                      |              |              |                 |               |              |
|------------|--------------------------------------|--------------|--------------|-----------------|---------------|--------------|
| <b>FP</b>  | ATLAS with direct Find-box method    | 879/880      | <b>0.913</b> | <b>0.927 mm</b> | <b>1.244%</b> | <b>91.2%</b> |
| <b>FP</b>  | ATLAS with segmentation-based method | 879/880      | <b>0.913</b> | 0.953 mm        | 1.508%        | 88.1%        |
| <b>UCL</b> | Median measurement                   | N/A          | N/A          | 26.400 mm       | 29.777%       | 5.5%         |
| <b>UCL</b> | Centred median box                   | N/A          | N/A          | 15.252 mm       | 21.170%       | 1.8%         |
| <b>UCL</b> | ATLAS with direct Find-box method    | <b>55/55</b> | <b>0.884</b> | 1.936 mm        | 2.843%        | 61.8%        |
| <b>UCL</b> | ATLAS with segmentation-based method | <b>55/55</b> | <b>0.884</b> | <b>1.432</b>    | <b>2.438%</b> | <b>79.6%</b> |

## Adaptation and Aggregation

Figure 4 presents the performance of Find and Measure across site-specific adapted and aggregated models.

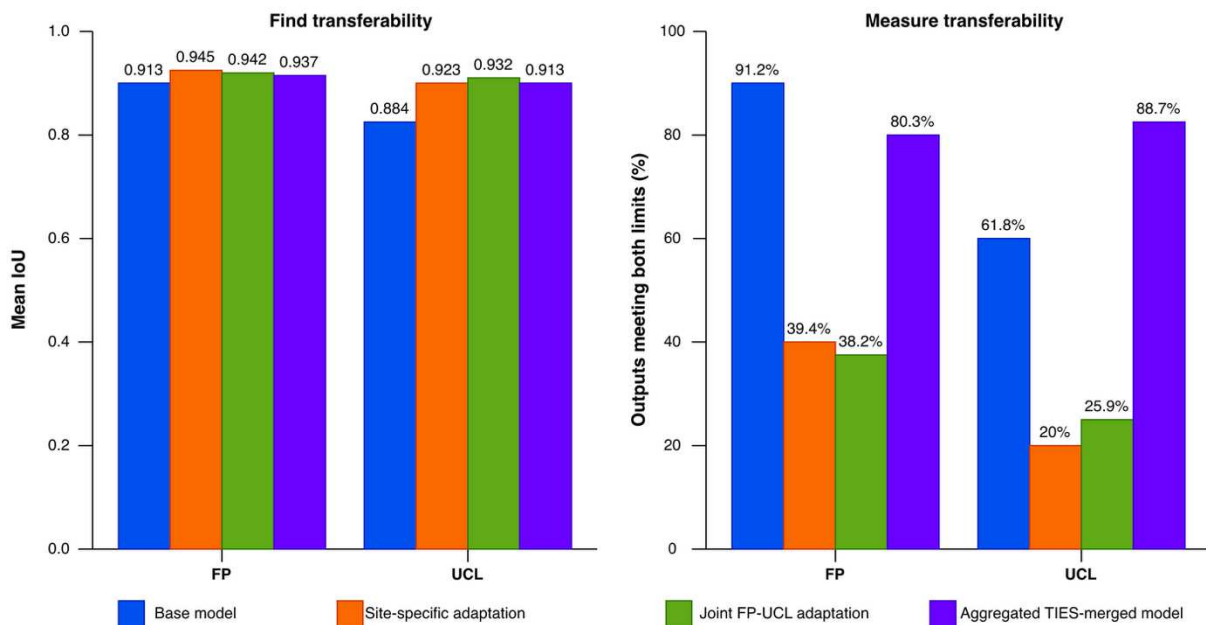


Figure 4. Find's localization and Measure's biometric performance across adapted and aggregated models.

### Site-specific adaptation

Site-specific adaptation was applied only to the Find detector; Confirm and segmentation-based Measure were not retrained. On FP, the specialist model increased mean cranial-box IoU from 0.913 to 0.945. However, BPD MAE increased from 0.927 to 2.628 mm, HC error increased from 1.244% to 3.546%, and the proportion meeting both accuracy limits decreased from 91.2% to 39.4%. The

1032 joint FP–UCL model exhibited a similar mismatch, with an IoU of 0.942 but only 38.2% of  
1033 available outputs meeting both limits.

1034 On UCL, site-specific adaptation increased IoU from 0.884 to 0.923. OFD error decreased from  
1035 3.716 to 2.943 mm, but BPD error increased from 1.936 to 3.384 mm, HC error increased from  
1036 2.843% to 3.168%, and the combined-limit rate decreased from 61.8% to 20.0%. The pooled model  
1037 achieved the highest UCL IoU at 0.932, but only 25.9% of outputs met both limits. These results  
1038 indicate that improved overall box overlap did not reliably preserve the dimensions required for  
1039 fetal-head biometry.

1040 Site-specific adaptation improved the overlap between predicted and reference cranial boxes;  
1041 however, the measurements calculated from those boxes became less accurate. On FP, fine-tuning  
1042 increased mean cranial-box IoU from 0.913 to 0.945. However, BPD MAE increased from 0.927  
1043 mm to 2.628 mm, HC error increased from 1.244% to 3.546%, and the proportion of outputs  
1044 meeting both accuracy limits decreased from 91.2% to 39.4%. Therefore, the base Find model  
1045 produced more accurate Find-derived measurements on FP than the FP-specific model. A similar  
1046 pattern was observed on UCL. Site-specific fine-tuning increased mean IoU from 0.884 to 0.923,  
1047 but BPD error increased from 1.936 mm to 3.384 mm, and the proportion meeting both accuracy  
1048 limits decreased from 61.8% to 20.0%. Again, the base model produced more accurate Find-derived  
1049 measurements than the site-specific model.

1050 These results demonstrate that greater overall box overlap does not guarantee more accurate fetal  
1051 biometry. A predicted box may achieve higher IoU while its horizontal or vertical boundaries shift  
1052 sufficiently to alter the BPD, OFD, and HC calculated from its dimensions. For this task, adapted  
1053 models should therefore be selected based on biometric accuracy rather than localization overlap  
1054 alone.

1055 **Aggregation of site-adapted models**

1056 TIES-Merging combined the FP- and UCL-adapted Find models using 2,523,182 aligned floating-  
1057 point checkpoint elements. At the prespecified density of 0.20, raw sign interference between the  
1058 site task vectors was 46.27%. Trim reduced interference to 9.14%, and Elect Sign eliminated the  
1059 remaining interference. The merged update contained 839,070 nonzero elements, and the  
1060 reconstructed mode; loaded and executed without missing or unexpected tensor keys. Post-Elect-  
1061 Sign interference remained zero throughout the density sensitivity analysis from 0.05 to 0.40.

1062 Eliminating parameter conflict did not result in uniform biometric improvement. On FP, the TIES  
1063 model achieved a BPD MAE of 1.386 mm, HC error of 2.137%, and 80.3% of available outputs  
1064 meeting both limits. This model recovered much of the performance lost due to site-specific and  
1065 pooled adaptation, but remained below that of the unchanged HC18 model. On UCL, TIES achieved  
1066 the lowest BPD and HC errors among the evaluated Find models, at 1.167 mm and 1.651%,  
1067 respectively, with 88.7% of available outputs meeting both limits. However, Find output was  
1068 available for 53 of 55 frames, compared with all 55 for the unchanged and UCL-specialist models.  
1069 TIES therefore provided the strongest overall result among the adapted models, in particular for  
1070 UCL cohort.

1071 **Computational performance and model footprint**

1072 ATLAS remained lightweight despite integrating three distinct analytical stages. The heavier  
1073 pathway Find, Confirm, and segmentation-based Measure models occupied 5.12 MB, 5.16 MB, and  
1074 6.22 MB, respectively, resulting in a total active model footprint of 16.50 MB.

1075 Mean workstation processing times were 14.04 ms for Find, 25.58 ms for Confirm, and 17.70 ms  
1076 for segmentation-based Measure, resulting in a total mean inference time of 57.32 ms per frame.  
1077 These results indicate that the model components are computationally compact. It should be noted  
1078 that these measurements do not include image transfer, interface rendering, device integration,  
1079 sustained-load performance, thermal behavior, or cybersecurity overhead.

## 1080 **Comparison with Prior Works**

1081 Previous studies on fetal-ultrasound AI have focused on various components of the biometry  
1082 workflow. The comparison with ATLAS is structured along two complementary axes: measurement  
1083 performance, where endpoints are sufficiently similar to allow for numerical comparison (Table 9),  
1084 and evidentiary and functional scope, where related approaches are considered but direct numerical  
1085 comparison would be misleading (Additional File 1).

1086 Most fetal-head measurement methods employ a segmentation-to-ellipse approach, in which the  
1087 skull is segmented, the contour is post-processed, and an ellipse is fitted to obtain HC and, in some  
1088 cases, BPD and OFD. Examples include DAG V-Net, the lightweight CSM model, SAP-UNet, and  
1089 TransFSM (43,44,56,57). Collectively, these studies demonstrate that increasingly sophisticated  
1090 attention-based, multi-scale, boundary-aware, and lightweight segmentation architectures can  
1091 produce accurate fetal-head contours and HC estimates. However, the principal evidence from these  
1092 studies concerns segmentation and measurement agreement after an appropriate fetal-head image  
1093 has entered the measurement pathway. The segmentation-based Measure method in ATLAS  
1094 achieved an HC error of 2.792 mm (1.242%) and a BPD MAE of 1.069 mm on the locked internal  
1095 test set, placing it within the range reported for this class of method (Table 9).

1096 Some studies have questioned the necessity of segmentation in all cases. For example, Zhang et al.  
1097 compared segmentation-based ellipse fitting with direct image-to-HC regression, finding that the  
1098 best segmentation model produced lower HC error, while regression required less inference time.  
1099 This comparison establishes an important precedent for evaluating alternative approaches to the  
1100 same biomarker. ATLAS addresses a related but distinct question. Its direct Find-box measurement  
1101 method is not an end-to-end regression model; instead, it calculates BPD, OFD, and HC  
1102 deterministically from geometry already produced by the Find stage. The comparison therefore  
1103 evaluates whether an additional learned segmentation stage improves the final measurement beyond  
1104 an established localization result. Internally, this was not the case: direct geometry achieved a BPD  
1105 MAE of 0.779 mm compared to 1.069 mm for segmentation, with a higher joint-tolerance rate

1106 (91.6% versus 84.1%). Externally, neither method consistently outperformed the other, with direct  
1107 geometry performing better on FP (0.927 versus 0.953 mm) and segmentation performing better on  
1108 UCL (1.432 versus 1.936 mm). These results indicate that the optimal geometry source depends on  
1109 the acquisition and annotation characteristics of the target domain.

1110 Staged fetal-ultrasound systems were developed prior to ATLAS. For instance, Rasheed et al.  
1111 classified fetal-head frames from ultrasound video before segmentation and measurement, and  
1112 Hekal et al. employed YOLOv5 localization prior to SAP-UNet segmentation and ellipse fitting.  
1113 SonoNet and LPC-SonoNet address the earlier challenge of detecting predefined fetal standard  
1114 views, with SonoNet additionally localizing anatomy using weak image-level supervision. Thus,  
1115 ATLAS is not the first system to integrate localization, frame selection, or multiple computational  
1116 stages. Its distinction lies in introducing a separate anatomical-evidence function and retaining the  
1117 evidence, scale source, geometry, and availability state associated with each individual  
1118 measurement. This approach enables the validity and traceability of a fetal-head measurement to be  
1119 represented and audited, including the assessment of whether alternative geometry sources and  
1120 model adaptations preserve the intended clinical quantity.

1121 Recent research has addressed limitations related to this contribution. Singh et al. combined fetal-  
1122 head segmentation with Grad-CAM++ to provide post hoc visualization of model attention, while  
1123 ERSR utilized ellipse-constrained pseudo-label refinement and consistency regularization to reduce  
1124 reliance on expert segmentation annotations. These approaches enhance explainability and  
1125 annotation efficiency, respectively. In contrast, ATLAS represents selected anatomical evidence  
1126 and provenance as structured fields within the measurement pathway, rather than visualizing model  
1127 attention after segmentation or improving the quality of training masks.

1128 The computational literature demonstrates a parallel trend toward lightweight models suitable for  
1129 site-specific and privacy-preserving deployment. The CSM fetal-head model contains 0.13 million  
1130 parameters with a reported 28 ms CPU inference time, while RTSeg-Net contains 1.86 million

parameters and achieves 31.13 frames per second on a Jetson Nano for intrapartum fetal-head and pubic-symphysis segmentation. ATLAS occupies 16.50 MB across all three stages, with a total workstation latency of 57.32 ms per frame. Although these figures are broadly similar, they are not directly comparable: the comparators are single-task segmentation models benchmarked on embedded hardware, whereas ATLAS reports a complete localization, evidence, and measurement pathway measured on a workstation. Establishing embedded-device performance for ATLAS would require equivalent benchmarking on comparable hardware.

In summary, the individual components of ATLAS, localization, segmentation, direct measurement, weak supervision, explainability, and lightweight implementation, each have established precedents. However, there is no prior work identified that evaluates these elements within a single pathway that also incorporates patient-audited data preparation, selected internal anatomical evidence, two alternative measurement geometries, per-output scale and provenance records, an explicit output-availability state, independent external transfer, and measurement-aware evaluation of site adaptation and model aggregation.

Table 9. Representative fetal-head studies reporting HC measurement error

| Method                              | Evaluation protocol reported by study                                             | Measurement pathway                           | HC absolute error        | Other reported results            | Efficiency or footprint                               |
|-------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------|--------------------------|-----------------------------------|-------------------------------------------------------|
| <b>DAG V-Net (44)</b>               | HC18 challenge test, n=355 as reported                                            | Head segmentation → post-processing → ellipse | ADF<br>1.77 ±<br>1.69 mm | Dice 97.93%;<br>HD 1.29 ± 0.79 mm | 6.3 ms/image on study GPU workstation                 |
| <b>U-Net–ResNet50 (58)</b>          | HC18; random 600/199/200 train/validation/test splits, averaged across five folds | Segmentation → post-processing → ellipse      | MAE<br>1.08 ±<br>1.25 mm | Dice 98.8 ± 0.9%                  | 68.26 s/200 images on CPU, approximately 0.34 s/image |
| <b>EfficientNet regression (58)</b> | Same custom HC18 protocol                                                         | Direct image-to-HC regression; no contour     | MAE<br>1.83 ±<br>2.11 mm | No segmentation output            | 36.95 s/200 images on CPU, approximately 0.18 s/image |
| <b>Lightweight CSM (56)</b>         | HC18, 999 training and 335 test images as reported                                | Lightweight segmentation → ellipse            | ADF<br>1.97 ±<br>1.89 mm | Dice 97.61 ± 1.72%                | 0.13 M parameters; 28 ms CPU and 0.194 ms GPU         |
| <b>SAP-UNet with YOLOv5 (43)</b>    | Exam-exclusive random 80/20 split of                                              | Head detection → segmentation → ellipse       | ADF<br>1.11 ±<br>1.30 mm | Dice 98.02 ± 0.80%; HD            | NR                                                    |

|                                                       |                                                                                                                              |                                                                          |                       |                                                                      |                                                              |
|-------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------|-----------------------|----------------------------------------------------------------------|--------------------------------------------------------------|
|                                                       | the 999-image HC18 training resource                                                                                         |                                                                          |                       | 1.27 ± 1.40 mm                                                       |                                                              |
| <b>TransFSM (57)</b>                                  | Custom HC18 split: 800 training and 199 test images                                                                          | Multi-anatomy segmentation → fitted measurement geometry                 | MAE 1.19 mm           | Dice 98.11%; HD 1.03 mm; signed difference 0.53 mm                   | Teacher footprint NR; distilled model had different accuracy |
| <b>ATLAS—segmentation-based measurement method</b>    | Patient-exclusive HC18 test: 586 frames from 279 patients<br>External test: FP and UCL<br>Same locked patient-exclusive test | Find → Confirm → head segmentation → fitted ellipse                      | MAE 2.792 mm (1.242%) | BPD MAE 1.069 mm; OFD MAE 1.510 mm; 84.1% met both BPD and HC limits | Complete three-stage profile: 16.50 MB and 57.32 ms/frame    |
| <b>ATLAS—direct Find-box-based measurement method</b> |                                                                                                                              | Find → Confirm → Measure (direct deterministic geometry or segmentation) | MAE 2.830 mm (1.259%) | BPD MAE 0.779 mm; OFD MAE 1.484 mm; 91.6% met both limits            | Complete three-stage profile: 10.28 MB and 39.62 ms/frame    |

**Note:** HC ADF and MAE both represent absolute HC disagreement in millimeters but were calculated under different study protocols. The table must not be interpreted as a rank ordering. Studies used different HC18 partitions, patient or image splitting strategies, reference-processing methods and statistical summaries. ATLAS's HC percentage is the mean absolute percentage error. Runtime values were measured on different hardware and are not directly comparable. NR, not reported.

## Clinical expert quality review

Comparison of ATLAS-FN test outputs with the MFM fellow's clinician-labelled assessment of the corresponding images showed that the stage-specific outputs were generally interpretable in relation to clinically recognizable features of fetal-head image quality and biometric suitability. The review identified both concordant cases, in which cranial localization, anatomical evidence, and measurement geometry were consistent with the clinician's assessment, and discordant cases in which one computational stage remained technically available despite limitations in the image for standardized biometry. These observations supported the intended separation of Find, Confirm, and Measure as related but distinct evidence states rather than as a single determination of clinical plane adequacy.

A particular observation concerned non-standard cranial sections. The MFM fellow noted that ATLAS retained interpretable cranial localization and geometry in several images at or near the transcerebellar level, including cases in which acquisition of the preferred biometric plane was difficult. Although such images would not necessarily be selected for standardized BPD/HC measurement, ATLAS could still preserve identifiable cranial and anatomical information. The

1166 review therefore highlighted that successful localization or geometric measurement availability  
1167 should not, by itself, be interpreted as evidence that the image satisfies all requirements for  
1168 standardized fetal-head biometry.

1169 The clinical review also indicated that images considered suboptimal for the specific task of  
1170 standardized biometry were not necessarily devoid of clinically interpretable information. Senior  
1171 MFM review of representative outputs and observed patterns supported this interpretation and the  
1172 positioning of ATLAS as an evidence-traceability and quality-assurance framework rather than an  
1173 autonomous determination of complete clinical plane adequacy.

1174

## 1175 **Discussion**

### 1176 **Interpretation of the key findings**

1177 This study developed and evaluated ATLAS, a staged framework that incorporates a data integrity  
1178 layer for patient-exclusive development and evaluation, ensuring traceability and accountability; a  
1179 Find component that localizes the fetal cranial region, a Confirm component that records whether  
1180 critical internal anatomy, specifically the CSP and LV, is visible within that region according to the  
1181 ISUOG standard, and a Measure component that derives geometric measurements of BPD, OFD,  
1182 and HC. ATLAS addresses a longstanding challenge in clinical measurement AI: silent,  
1183 unexplained error. In fetal biometry, selecting an inappropriate imaging plane can introduce  
1184 measurement discrepancies of several millimeters, often exceeding the variation associated with  
1185 caliper placement alone, which undermines accuracy, reliability, and clinical trust. ATLAS shifts  
1186 the unit of accountability in medical measurement AI from the output to the entire measurement  
1187 pathway.

1188 Six interconnected findings emerged from this study. First, the data-integrity checks were  
1189 complementary and essential in AI medical imaging; each identified a distinct class of data defect,  
1190 thereby enabling responsible and more reliable AI results with traceable evidence (59). Second, the

1191 Find-Confirm-Measure pathway contributed distinct information, allowing each measurement to be  
1192 traced back to its localization and the supporting anatomical evidence. Third, the two measurement  
1193 methods, deterministic geometry from the Find box and geometry fitted to a learned segmentation  
1194 contour, were complementary and suitable for different ultrasound data characteristics. Fourth,  
1195 ATLAS models demonstrated transferability for localization across nearly all external dataset  
1196 frames, although measurement accuracy varied depending on external imaging and scale  
1197 conditions. Fifth, site-specific adaptation improved localization but reduced measurement accuracy,  
1198 a loss that aggregation partially recovered, particularly in the UCL test set. Finally, the pathway  
1199 remained computationally lightweight, at 16.50 MB and 57.32 ms per frame, supporting future  
1200 cross-site edge computing integration.

1201 ATLAS occupies an intermediate position between narrow single-task models and whole-  
1202 examination automation. Existing single-task models address standard-plane detection, anatomical  
1203 analysis, and biometry estimation as separate tasks (41), while end-to-end whole-scan AI systems  
1204 automate larger segments of the workflow (60,61). Rather than replacing these approaches, ATLAS  
1205 bridges the gap between accuracy-focused single-task models and whole-scan automated systems,  
1206 facilitating clinical user involvement. It builds upon the concept that explainable AI-enabled fetal  
1207 ultrasound models should employ stagewise layers aligned with clinician reasoning, and that plane  
1208 identification should be supported by key CSP and LV evidence (62,63). ATLAS is distinct in that  
1209 it links questions typically evaluated separately: whether the fetal head was found, what anatomical  
1210 evidence was detected, how measurement geometry was generated, how pixels were converted into  
1211 millimeters, whether an output was available, and whether adaptation preserved the final biometric  
1212 value. The unit of evaluation is not a box, contour, landmark, or number, but the traceable pathway  
1213 through which the measurement was produced. This distinction is clinically significant because an  
1214 oblique or incorrectly positioned section can yield a plausible value while systematically  
1215 misestimating fetal size, with downstream consequences for growth centiles and decisions  
1216 regarding repeat scanning, referral, or surveillance. ATLAS does not evaluate these clinical

1217 decisions; rather, it addresses whether the basis of an automated measurement is sufficiently visible  
1218 for review.

1219 The data-integrity results establish data preparation as an integral part of the evidence chain, rather  
1220 than as routine preprocessing. Patient-exclusive partitioning, which assigns all frames from a single  
1221 patient to one partition, is essential because a single fetus or examination may contribute several  
1222 closely related frames. If such frames cross partitions, apparently strong AI model performance  
1223 may reflect recognition of related anatomy rather than true generalization to unseen patients, the  
1224 common barrier from laboratory-developed AI system to real practice. Controlled data fault-  
1225 injection testing, in which known defects were deliberately introduced, demonstrated that the full  
1226 integrity layer detected cross-partition patient overlap, duplicated frame identifiers, missing labels,  
1227 invalid pixel spacing, and discrepancies between observed and expected cohort counts. These  
1228 findings support existing arguments that ethical medical-imaging AI must address training data  
1229 leakage and the associated privacy, data quality, fairness, and transparency risks (59,64). The  
1230 integrity layer maintains the persistence of the link between each output and its source image,  
1231 patient partition, annotation origin, pixel scale, measurement method, and availability status,  
1232 ensuring that the evidence supporting a measurement can be examined rather than assumed to be  
1233 uniformly reliable.

1234 The Find–Confirm–Measure sequence mirrors clinical reasoning, wherein the operator acquires the  
1235 cranial plane, verifies internal anatomy, and then places or reviews the calipers (62). Frames in  
1236 which the Confirm component detected CSP evidence yielded more favorable measurements with  
1237 both methods, particularly for BPD derived from the Find box. This suggests that the anatomical-  
1238 evidence record provides information relevant to interpreting the measurement. However,  
1239 segmentation-based measurement results are variable. Therefore, the Confirm component is best  
1240 utilized as supplementary information for review. These findings support and provide insight into

1241 existing challenges in landmark-aware brain-plane systems, where CSP and LV visibility are  
1242 clinically meaningful but can be difficult to detect in certain abnormalities (63).

1243 A comparison of the two measurement methods indicates that both perform well depending on the  
1244 data environment and characteristics. Measurement results diverge due to differences in three key  
1245 dataset axes: reference construction (HC18's ellipse-derived labels favor the axis-aligned box,  
1246 which closely reproduces ellipse axes at  $r \approx 0.999$ , whereas FP's and UCL's manual calipers favor a  
1247 fitted contour); in-plane head orientation and framing (tightly controlled in HC18, highly variable  
1248 in UCL, which challenges the box's axis-aligned assumption); and image and scale heterogeneity  
1249 (clean, single-vendor HC18/FP data reward the box's simplicity, while UCL's multi-operator, partly  
1250 fallback-scaled frames reward the ellipse's rotational robustness). Segmentation systematically  
1251 under-measures BPD (signed error  $-0.84$  mm versus  $-0.15$  mm for direct), which is a disadvantage  
1252 on clean data but negligible where framing error dominates. Consequently, the box method is  
1253 optimal for HC18 and FP, while the ellipse method is preferable for UCL. Direct Find-box geometry  
1254 is recommended when acquisition is clean, the plane is well-framed and near axis-aligned, and the  
1255 reference or validation convention is ellipse- or box-based, as in single-vendor, protocol-controlled  
1256 data such as HC18 and, to a large extent, FP. Segmentation-based measurement is preferable when  
1257 acquisition is heterogeneous, framing and in-plane head orientation vary, and the reference follows  
1258 manual caliper placement, as in multi-site, multi-operator data such as UCL. The crossover point  
1259 depends on whether the axis-aligned box remains a valid proxy for the true BPD/OFD axes; as in-  
1260 plane rotation and framing variability increase, the rotatable fitted ellipse becomes superior. This  
1261 challenges the prevailing notion that treats segmentation as merely an intermediate step to ellipse  
1262 fitting, demonstrating that direct estimation approaches can be competitive and that richer  
1263 representations do not always yield better measurements (41,58).

1264 External evaluation separated two outcomes that are often reported together: the proportion of  
1265 frames for which a cranial region was detected, and the correctness of the measurement derived

1266 from that region. Find returned a usable region in 99.9% of FP frames and in all UCL frames, yet  
1267 Find-box measurements were less accurate on UCL. A model can therefore identify the correct  
1268 general region while producing dimensions that are less reliable. Pixel-spacing provenance  
1269 compounded this, as 36.4% of UCL frames required fallback spacing rather than image-specific  
1270 metadata. When scale information is unavailable, the anatomical location may remain reasonable  
1271 in pixel coordinates while the final value in millimeters becomes less certain. This finding informs  
1272 the need for any evaluation to consider how physical scale was obtained, not only whether an output  
1273 was generated, a requirement that extends to any imaging application in which model-generated  
1274 geometry must be converted into clinically interpreted units.

1275 The site-adaptation experiment yielded a consequential finding. Fine-tuning the model on each  
1276 external cohort improved the IoU between predicted and reference boxes. Although the adapted  
1277 models more closely enclosed the fetal head overall, individual box boundaries shifted sufficiently  
1278 to distort the dimensions used to calculate BPD, OFD, and HC. Therefore, adapted models should  
1279 not be selected or approved based solely on localization performance; they must also be assessed  
1280 for biometric error, bias, output availability, and performance across relevant patient and device  
1281 groups prior to deployment. This finding supports existing distributed model learning literature,  
1282 which demonstrates that optimizing for a single objective or site does not guarantee improved cross-  
1283 site task performance (65).

1284 In contrast, model aggregation recovered more measurement performance than either site-specific  
1285 or pooled adaptation, although the benefit was dataset dependent. On FP, aggregation restored much  
1286 of the loss introduced by adaptation, though it did not reach the performance of the unadapted  
1287 HC18-trained model. On UCL, aggregation produced the strongest measurement performance, with  
1288 88.7% of available outputs meeting both limits, but failed to return a localization for two frames.  
1289 This finding is consistent with literature indicating that aggregation can enhance both site  
1290 specialization and generalization (52). Combining site-adapted parameters without pooling source

1291 images during merging is therefore feasible, and evaluating the merged model based on resulting  
1292 biometry, rather than parameter compatibility or localization accuracy alone, advances current  
1293 practice in distributed medical AI. This image-free merging approach extends the federated learning  
1294 literature on collaborative learning without image sharing in fetal ultrasound (51,66) and  
1295 demonstrates the potential for privacy-preserving multi-institutional imaging AI in this field (64).

1296 The clinical expert review provided complementary evidence that the evidence-separated design of  
1297 ATLAS was clinically interpretable. Localization, anatomical evidence, and biometric suitability  
1298 could diverge in clinically meaningful ways. This was apparent in some non-standard and  
1299 transcerebellar-level views, where ATLAS retained interpretable cranial localization and geometry  
1300 although the image would not necessarily be selected for standardized BPD/HC measurement. This  
1301 observation does not validate biometry from the transcerebellar plane; rather, it reinforces that  
1302 computational availability is not equivalent to fitness for a particular clinical measurement task.  
1303 Conversely, failure to satisfy the requirements of standardized biometry does not necessarily render  
1304 an image devoid of other clinically interpretable information. This finding is consistent with  
1305 emerging medical-imaging literature showing that AI confidence may remain high when evidence  
1306 is insufficient, that ultrasound quality is task-dependent, and that uncertainty- or feature-specific  
1307 assessments can support graded rather than binary handling of images (67,68).

1308 ATLAS extends these approaches by retaining localization, selected anatomical evidence,  
1309 measurement geometry, scale provenance, and output availability as linked but distinct evidence  
1310 states associated with the individual measurement. This differs from approaches that primarily  
1311 identify or filter optimal images before downstream analysis, where information from non-optimal  
1312 acquisitions may be discarded rather than qualified according to its remaining evidential value  
1313 (69,70). In this sense, the framework moves from image-level quality assessment toward task-  
1314 specific evidence assurance: the relevant question becomes not simply whether an image or output  
1315 is “valid,” but what the available evidence is sufficient to support. Making that evidence inspectable

1316 also creates a basis for more verifiable human–AI interaction, because clinicians can interrogate the  
1317 evidence underlying an automated result rather than relying on the numerical output or confidence  
1318 score alone. This provides a foundation for future research into which forms of evidence clinicians  
1319 actually use to accept, question, reacquire, or disregard AI-generated measurements, how such  
1320 evidence affects appropriate reliance and workflow integration, and how clinical feedback could in  
1321 turn inform the design of quality-aware AI systems.

## 1322 **Implications**

1323 For biomedical informatics, these findings suggest that quality assurance should be represented  
1324 within the information architecture of an AI system rather than reduced to a confidence score or  
1325 final pass/fail decision. Existing work has shown the value of uncertainty-aware prediction, quality  
1326 scoring, feature-level reasoning and expert review for identifying outputs that require caution  
1327 (71,72). ATLAS builds on this direction by preserving an auditable chain from image eligibility  
1328 and cranial localization to the resulting biometric value. This changes the role of quality control  
1329 from output rejection to evidence-preserving qualification: unsupported downstream uses can be  
1330 withheld or questioned without discarding information that remains valid for another purpose. Such  
1331 a record may also provide a common review object across clinicians, developers, researchers and  
1332 governance teams, allowing a questionable measurement to be traced to a specific failure or  
1333 uncertainty state rather than explained only through post-hoc visualization.

1334 While plane-aware commercial systems may have undergone regulatory evaluation for defined  
1335 intended uses, regulatory clearance does not necessarily provide clinicians with a transparent, per-  
1336 output evidence chain that can be independently interrogated or audited (19,73). This creates an  
1337 important integration gap: clinicians may receive an automated result without sufficiently granular,  
1338 verifiable information to assess its evidential basis, appropriately calibrate reliance, or identify the  
1339 stage at which an error may have arisen. Large Language Model-layered systems may offer  
1340 persuasive rationales for predictions but do not supply users with the verifiable data underlying  
1341 those predictions (74–76). In contrast, ATLAS's per-output record provides clinical reviewers and

1342 governance teams with a common, verifiable object, enabling audit at the level of individual  
1343 measurements, which is more granular than system-wide clearance.

1344 Two additional implications relate to data management and deployment. The integrity layer  
1345 provides a reproducible solution to a common challenge in clinical imaging-AI development, where  
1346 expert annotations, derived targets, metadata, and computationally generated labels are frequently  
1347 treated as equivalent. By recording patient identity, annotation origin, negative-label status,  
1348 transformation history, pixel-spacing source, checkpoint identity, and output within a single ledger,  
1349 clear boundaries are maintained regarding what has and has not been clinically annotated. Confirm  
1350 demonstrates that lightweight models can be trained within this framework using responsible weak  
1351 supervision when expert masks are unavailable. Furthermore, the 16.50 MB model size and 57.32  
1352 ms runtime indicate that incorporating explicit eligibility and evidence stages does not necessitate  
1353 a large foundation model. This makes a quality-conditioned, evidence-linked pipeline feasible at  
1354 the point of care, where latency, connectivity, and privacy constraints often preclude console-bound  
1355 or cloud-dependent inference. When combined with data-free aggregation, which merges site-  
1356 adapted parameters without transferring raw images, this approach supports cross-institutional and  
1357 low-resource deployment. It also promotes equitable access, as most obstetric scanning globally  
1358 lacks the financial and infrastructural resources to support advanced AI platforms.

1359 From a clinical perspective, automated biometry should provide users with sufficient information  
1360 to understand how a measurement was obtained and to determine whether to accept, inspect,  
1361 reacquire, or disregard it. BPD or HC values may appear convincing even if derived from an oblique  
1362 section, a mislocalized region, unsuitable geometry, or uncertain scaling. ATLAS addresses this by  
1363 presenting a single review object that allows users to evaluate each component, supported by a dual-  
1364 context view that magnifies internal structures while preserving acquisition context. Further review  
1365 on the ATLAS results further suggested that preserving this wider context may retain clinically  
1366 relevant information even when an image is suboptimal for standardized biometry; for example,

1367 wider or inadequately magnified views may still assist clinical interpretation when considering  
1368 suspected small- or large-for-gestational-age (SGA/LGA) presentations. This does not imply that  
1369 ATLAS identifies fetal growth abnormalities, but illustrates why task-specific measurement  
1370 adequacy should not be equated with the absence of other clinically useful image information.  
1371 Overall, this approach may foster more appropriate reliance and contextual use, particularly under  
1372 time constraints when automation bias can lead to acceptance of plausible but incorrect results. The  
1373 adaptation findings further inform model governance; for example, an update that enhances  
1374 localization may introduce systematic changes in BPD or HC measurements, which is more  
1375 problematic than random variation as it consistently shifts results. Consequently, updates should be  
1376 managed as modifications to a clinical measurement process, with regression testing for bias,  
1377 availability, and subgroup performance compared to the deployed version.

1378 Review on ATLAS results on the ultrasound frames further informs why anatomical-evidence  
1379 detection and overall image adequacy should remain separate concepts. Selected internal structures  
1380 may remain detectable in images that do not satisfy another component of the ISUOG quality  
1381 framework—for example, when the acquisition level is suboptimal, the head is inadequately  
1382 positioned or magnified, or exclusionary anatomy is visible. Review by an MFM fellow also  
1383 highlighted a related clinical nuance: an image that is suboptimal for standardized head biometry  
1384 may nevertheless retain contextual information of potential clinical interest. For example, in some  
1385 wider-field acquisitions in which the fetal head appeared relatively high or insufficiently magnified  
1386 within the image, the surrounding fetal context could still be informative when considering fetuses  
1387 at the extremes of growth, such as suspected small- or large-for-gestational-age presentations. This  
1388 observation does not imply that such images satisfy biometric-plane requirements or that ATLAS  
1389 can identify fetal growth abnormalities; rather, it illustrates that technical adequacy for a specific  
1390 measurement task and the broader clinical information contained within an ultrasound image are  
1391 not equivalent. In such cases, successful CSP/LV detection indicates that the model has identified  
1392 the targeted structure, but does not convert the image into an acceptable biometric plane. This

1393 distinction supports the modular design of ATLAS: individual evidence states can be retained  
1394 explicitly and subsequently combined with additional criteria, rather than treating a single detected  
1395 landmark as a surrogate for overall clinical adequacy.

1396 In summary, ATLAS is designed to facilitate a more transparent fetal-head measurement workflow  
1397 by making the process visible, preserving the conditions under which each output is generated, and  
1398 establishing the technical foundation for quality-assurance and decision-support studies that will  
1399 determine its clinical value.

#### 1400 **Scope of inference**

1401 The scope of inference in this study is constrained by the acquisition characteristics of the HC18,  
1402 FP, and UCL datasets, the derivation of their reference standards, and the static-frame evaluation  
1403 design. All development and internal testing were conducted using HC18, a single-source repository  
1404 of pre-selected two-dimensional fetal-head planes acquired on a GE Voluson E8, consisting of 800  
1405  $\times$  540-pixel images with pixel sizes ranging from 0.052 to 0.326 mm. The external FP and UCL  
1406 cohorts were also acquired on GE Voluson systems. Therefore, while FP and UCL offer  
1407 independent subjects, sites, and annotation teams, the external evaluation is more accurately  
1408 characterized as cross-site and cross-cohort rather than cross-vendor transfer. Ultrasound  
1409 manufacturers exhibit substantial differences in beamforming, speckle reduction, harmonic  
1410 imaging, and post-processing, all of which influence the edge appearance critical for cranial  
1411 localization and contour fitting. As a result, generalization to other vendors or handheld point-of-  
1412 care systems remains untested. Furthermore, the nearly eightfold variation in pixel sizes within  
1413 HC18 alone highlights the scale heterogeneity that any such transfer would need to address.

1414 The reference standards constrain interpretation as significantly as the images themselves. HC18  
1415 provides expert ellipse annotations of the fetal head, from which both the Find boxes and the 72-  
1416 point polygon segmentation targets were derived, and against which internal BPD, OFD, and HC  
1417 were evaluated. An axis-aligned bounding box of a well-oriented reference ellipse closely

1418 approximates the ellipse's minor and major axes, making the deterministic Find-box comparator  
1419 geometrically well matched to the internal reference standard. This alignment offers a plausible  
1420 alternative explanation for its superior internal BPD and joint-tolerance performance, which should  
1421 be considered alongside the interpretation that box geometry more accurately preserves the required  
1422 axes. Two observations support a label-geometry effect: the internal advantage of direct geometry  
1423 was greatest on HC18, whose reference is ellipse-derived, and diminished on FP and reversed on  
1424 UCL, where references were constructed from manual expert landmarks (77), the comparison  
1425 should be interpreted as evidence that the optimal measurement pathway depends on the annotation  
1426 convention and imaging environment in which it is validated, rather than as evidence that learned  
1427 segmentation is inherently inferior. Definitive resolution requires evaluation against direct expert  
1428 contour masks and caliper placements obtained independently of the geometry under test.  
1429 Additionally, no inter-observer or intra-observer repeatability estimates were available for these  
1430 annotations, so the irreducible measurement floor remains unknown. Reported errors of  
1431 approximately 0.8–1.1 mm for BPD may approach the magnitude of expert disagreement, and small  
1432 differences between pathways should be interpreted accordingly.

1433 The anatomical evidence function is subject to its own limitations. The Confirm reference was based  
1434 on a box-derived evidence ledger rather than independent expert standard-plane adjudication, so its  
1435 performance reflects consistency with a curated evidence representation rather than certification of  
1436 plane suitability. The presence of the CSP and LV is necessary but not sufficient to establish a  
1437 complete transthalamic or transventricular plane, which also requires visualization of the thalami,  
1438 a symmetrical midline falx, and an appropriate axial level, none of which were verified. The datasets  
1439 do not indicate which of these two planes a given frame represents, so evidence expectations could  
1440 not be conditioned on the intended view. Gestational age introduces an additional constraint: the  
1441 CSP becomes reliably visible from approximately 17 to 20 weeks, so a "not detected" response is  
1442 anatomically expected rather than indicative of poor quality in earlier gestations. The gestational  
1443 distribution of the evaluated frames was not available for stratification. Importantly, the CSP may

1444 also be genuinely absent due to clinically significant conditions, such as agenesis of the corpus  
1445 callosum and holoprosencephaly. These pathologies are rare and minimally represented in these  
1446 datasets, so the behavior of Confirm in precisely the circumstances where its output would be most  
1447 clinically relevant remains uncharacterized. Its outputs should not be interpreted as anomaly  
1448 detection.

1449 Population-level inference is similarly constrained. Publicly available fetal-biometry datasets  
1450 contain limited demographic or clinical covariate information, preventing formal fairness  
1451 assessment across maternal ethnicity, body habitus, socioeconomic status, parity, comorbidity,  
1452 singleton versus multiple pregnancy, or fetal growth status. As a result, performance in populations  
1453 where image quality is systematically more difficult to obtain, cannot be estimated from these data.  
1454 Measurement convention introduces an additional clinical limitation: the reference geometry in this  
1455 study follows the outer skull perimeter, consistent with ISUOG practice, whereas ASUM  
1456 recommends an outer-to-inner caliper placement for BPD. Therefore, the reported errors are  
1457 referenced to one convention and would require recalibration, in settings using the other.

#### 1458 **Strengths, limitations, and future research**

1459 The principal strength of this study lies in its evidence-linked, ablation-validated design, which  
1460 demonstrates that patient-exclusive data governance and explicit stage separation are necessary for  
1461 trustworthy measurement. The paired component comparison, conducted with patient-cluster  
1462 uncertainty over a locked test cohort, isolates the effect of each measurement pathway on the  
1463 clinical endpoint rather than on an intermediate vision metric. The controlled data-integrity ablation  
1464 demonstrates that each integrity check addresses a distinct threat to evaluation validity.  
1465 Furthermore, the external and aggregation analyses report negative and mixed findings:  
1466 segmentation was not superior for every endpoint, higher localization overlap did not preserve  
1467 biometry, and model merging benefited one domain while adversely affecting another, rather than  
1468 reporting only favourable metrics. Collectively, these findings strengthen the credibility of ATLAS  
1469 as an auditable measurement architecture rather than a single-metric model.

1470 Several limitations should be acknowledged. First, the framework was evaluated on retrospective  
1471 static-frame benchmarks that do not replicate live probe movement, repeated acquisition, off-plane  
1472 distractors, or the temporal context of a complete examination. Computational feasibility was  
1473 established on a workstation and did not include image transfer, rendering, ultrasound-vendor  
1474 integration, sustained-load behaviour, or cybersecurity, and therefore does not establish console or  
1475 edge-device readiness. Second, the Confirm and segmentation reference standards were derived  
1476 from human-curated boxes rather than independent expert masks, so the reported evidence and  
1477 contour metrics reflect internal consistency rather than adjudicated clinical accuracy. Fourth,  
1478 publicly available fetal-biometry datasets contain limited demographic and clinical covariate  
1479 information, which restricts formal fairness assessment across subgroups such as maternal habitus,  
1480 gestational age band, device vendor, and pathological anatomy. The near-absence of pathological  
1481 cases, such as agenesis of the corpus callosum where the CSP is absent for clinically critical reasons,  
1482 means the behavior of Confirm on such images remains uncharacterized. In addition, the datasets  
1483 did not provide sufficiently complete protected-attribute or sociodemographic metadata to assess  
1484 differential performance across patient groups; demographic fairness could therefore not be  
1485 evaluated. The external evaluation, therefore, should be interpreted as technical robustness rather  
1486 than as evidence of fairness.

1487 Several research priorities emerge. First, prospective, multi-center validation on routine clinical  
1488 practice is required before any clinical implementation. Such studies should include diverse  
1489 institutions, devices, operators, gestational ages, body habitus, and image quality conditions. They  
1490 should also test the unavailable-output state on genuinely off-plane and pathological acquisitions to  
1491 determine whether ATLAS abstains appropriately rather than measuring with undue confidence.  
1492 Second, a clinically complete extension of Confirm would require independent expert annotation  
1493 and prospective evaluation of the remaining ISUOG evidence dimensions, including thalamic  
1494 visualization, cerebellar exclusion, plane symmetry, magnification, and measurement placement,  
1495 where appropriate. This would allow the present selected-anatomy evidence model to evolve into a

1496 multidimensional image-quality evidence layer without assuming that any single landmark  
1497 establishes plane adequacy. Third, measurement-aware adaptation and aggregation objectives  
1498 should directly constrain the biometric endpoint, signed bias, and output availability. Any model  
1499 update should be governed as a change to a clinical measurement process, with regression testing  
1500 across devices and subgroups rather than approval based solely on localization performance. Fourth,  
1501 the structured stage outputs and the dual-context evidence record provide a technically tractable  
1502 foundation for human-centered deployment. Rather than presenting only a single value, future  
1503 implementations could provide an audit-ready summary that includes where the head was localized,  
1504 which anatomy was detected, which pixel-spacing source was used, and whether an output was  
1505 available. Recent advances in grounded, controlled natural-language generation offer a viable  
1506 pathway to bedside communication, provided the underlying evidence remains separately verifiable  
1507 [refs]. Ultimately, the definitive benchmark for clinical utility is not retrospective measurement  
1508 accuracy but whether deployment demonstrably improves the safety and reviewability of  
1509 intrapartum-equivalent decision-making. This includes reducing acceptance of measurements from  
1510 unsuitable images without increasing unnecessary reacquisition or alert fatigue. Establishing this  
1511 will require rigorously designed prospective and human-factors studies, potentially including  
1512 stepped-wedge or randomized designs, to evaluate how clinicians interact with the evidence record  
1513 and how it influences reliance, workload, and patient safety in real ultrasound workflows.

## 1514 **Conclusions**

1515 ATLAS offers a patient-audited, anatomically traceable framework for fetal-head ultrasound  
1516 biometry that addresses the multidimensional challenges in clinical imaging AI as it transitions from  
1517 research laboratories to clinical deployment, integration, and trust. The framework comprises four  
1518 components: a data integrity layer to overcome conversion barriers in clinical imaging datasets; the  
1519 Find stage for detecting fetal cranial structures; the Confirm stage for identifying anatomical  
1520 evidence; and the Measure stage for determining measurement geometry using both region-based  
1521 and segmentation-based methods.

1522 The system is designed to ensure that any generated value is reviewable based on the supporting  
1523 conditions, rather than being accepted as a self-sufficient number. External evaluation demonstrated  
1524 high localization performance with transferable measurement capability across previously unseen  
1525 multi-center datasets. These results illustrate how ATLAS shifts the unit of accountability in clinical  
1526 measurement AI from the output itself to the pathway that produces it. Traditionally, imaging AI  
1527 models have been viewed as functions mapping images to numbers, with human users certifying  
1528 the function. In contrast, ATLAS establishes a different paradigm: accountability resides in the  
1529 entire chain, eligibility, evidence, geometry, scale, availability, and provenance, and a measurement  
1530 is only as reliable as its weakest link. The findings indicate that a technically superior localization  
1531 score may coincide with substantially poorer biometry, that more complex models do not  
1532 necessarily outperform simpler geometric approaches, and that identical model weights can yield  
1533 different results across cohorts. These observations underscore the necessity of embedding  
1534 trustworthiness within the inference pathway and evaluating systems by the clinical quantities they  
1535 are intended to provide, rather than by convenient upstream proxies. This principle is broadly  
1536 applicable to any system converting model-derived geometry into clinical units. Regulatory trends  
1537 are moving toward lifecycle monitoring, transparency, and traceability for AI as a medical device;  
1538 ATLAS's per-output provenance provides one potential avenue toward that end. Furthermore, the  
1539 increasing resistance to opaque, overconfident generative systems in medicine is driving demand  
1540 for approaches that prioritize evidence before assertion and allow for abstention when appropriate.

1541 The enduring and broader value of ATLAS is best characterized as a transferable standard of proof  
1542 for clinical measurement AI. Its measurement-validity architecture, data integrity layer, and  
1543 demonstrated evaluation principles are model-independent. The framework's compact, data-  
1544 minimizing design may enable implementation at the point of care and across institutions that are  
1545 unable to pool imaging data. Future research should include prospective evaluation on live  
1546 scanning, particularly involving under-represented cases, as well as human-factors and workflow

1547 studies examining clinician interaction with the evidence record. Additionally, appropriate  
 1548 governance of model updates will be necessary as clinical measurement processes evolve.

1549

## 1550 **List of abbreviations**

| Abbreviation | Definition                                                                                                                |
|--------------|---------------------------------------------------------------------------------------------------------------------------|
| AI           | Artificial intelligence                                                                                                   |
| APPRAISE-AI  | APPRAISE-AI tool for quantitative evaluation of AI studies for clinical decision support                                  |
| ATLAS-FN     | Anatomically Traceable Localization, Assessment and Segmentation for Fetal Neurosonography                                |
| BPD          | Biparietal diameter                                                                                                       |
| CI           | Confidence interval                                                                                                       |
| CLAHE        | Contrast-limited adaptive histogram equalization                                                                          |
| CNN          | Convolutional neural network                                                                                              |
| CSP          | Cavum septi pellucidi                                                                                                     |
| DECIDE-AI    | Developmental and Exploratory Clinical Investigations of Decision-support systems driven by Artificial Intelligence       |
| FCM          | Find–Confirm–Measure                                                                                                      |
| FP           | Fetal Planes source/domain                                                                                                |
| HC           | Head circumference                                                                                                        |
| IoU          | Intersection over union                                                                                                   |
| ISUOG        | International Society of Ultrasound in Obstetrics and Gynecology                                                          |
| LGA          | Large for gestational age                                                                                                 |
| LV           | Lateral ventricle                                                                                                         |
| MAE          | Mean absolute error                                                                                                       |
| MFM          | Maternal–fetal medicine                                                                                                   |
| OFD          | Occipitofrontal diameter                                                                                                  |
| SAM          | Segment Anything Model                                                                                                    |
| SGA          | Small for gestational age                                                                                                 |
| STARD-AI     | Standards for Reporting Diagnostic Accuracy Studies–Artificial Intelligence                                               |
| TIES         | Trim–Elect Sign–Merge                                                                                                     |
| TRIPOD+AI    | Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence |
| UCL          | University College London Hospital source/domain                                                                          |
| YOLO         | You Only Look Once                                                                                                        |

## 1551 **Declarations**

### 1552 **Ethics approval and consent to participate**

1553 This study used de-identified retrospective public benchmark data and did not recruit participants  
 1554 or collect new patient data. No additional institutional ethics review was required for this  
 1555 retrospective evaluation.

1556 **Consent for publication**

1557 Not applicable.

1558 **Availability of data and materials**

1559 The development data used in this study were obtained from publicly available annotated fetal-  
1560 ultrasound resources available via Zenodo (DOI: 10.5281/zenodo.8265464). External evaluation  
1561 data from the FP and UCL cohorts were obtained from the publicly available *FetalBiometry-*  
1562 *MultiCentre-Landmarks* resource (DOI: 10.5522/04/30819911). Third-party imaging datasets are  
1563 not redistributed by the authors and should be obtained directly from their original repositories in  
1564 accordance with the applicable terms of use.

1565 The ATLAS software release is available through GitHub and is archived in Zenodo as ATLAS  
1566 v1.0.0 (DOI: 10.5281/zenodo.22155263). The release contains the manuscript-associated trained  
1567 model checkpoints and an executable Python inference notebook documenting checkpoint loading,  
1568 inference settings, required input structure, and generation of evidence-linked ATLAS outputs.  
1569 Availability and reuse of the released checkpoints remain subject to the applicable licensing terms  
1570 described in the repository.

1571 **Competing interests**

1572 The authors declare no competing interests.

1573 **Funding**

1574 This research received no specific grant from any funding agency in the public, commercial, or  
1575 not-for-profit sectors.

1576 **Authors' contributions**

1577 DT: conceptualization, methodology, implementation, evaluation, reproducibility, supervision, and  
1578 manuscript writing. EIY: result evaluation, experiment, reproducibility and data-integrity framing,  
1579 medical-informatics interpretation, and manuscript revision. FdSC: clinical domain expertise,  
1580 validation, interpretation, and manuscript review. All authors reviewed and approved the  
1581 manuscript.

1582 **Acknowledgements**

1583 The authors thank the organizers and contributors of the HC18, FETAL\_PLANES\_DB, FP, and  
1584 UCL public fetal-ultrasound resources, which made the development and methodological  
1585 evaluation of ATLAS-FN possible. The authors also thank Madeeha Dean, Maternal–Fetal  
1586 Medicine Fellow, for the independent clinical labelling and expert review of ATLAS test images  
1587 against ISUOG-relevant criteria for fetal-head biometry, and for providing clinical insights that  
1588 informed interpretation of the findings.

1589 **References**

- 1590 1. Salomon LJ, Alfirevic Z, Berghella V, Bilardo CM, Chalouhi GE, Da Silva Costa F, et al. ISUOG  
1591 Practice Guidelines (updated): performance of the routine mid-trimester fetal ultrasound scan.  
1592 *Ultrasound Obstet Gynecol.* 2022;59(6):840–56. doi:10.1002/uog.24888
- 1593 2. Malinger G, Paladini D, Haratz KK, Monteagudo A, Pilu GL, Timor-Tritsch IE. ISUOG Practice  
1594 Guidelines (updated): sonographic examination of the fetal central nervous system. Part 1: performance  
1595 of screening examination and indications for targeted neurosonography. *Ultrasound Obstet Gynecol.*  
1596 2020 Sep 1;56(3):476–84. doi:10.1002/uog.22145 PubMed PMID: 32870591.
- 1597 3. Yang X, Dou H, Huang R, Xue W, Huang Y, Qian J, et al. Agent With Warm Start and Adaptive  
1598 Dynamic Termination for Plane Localization in 3D Ultrasound. *IEEE Trans Med Imaging.* 2021  
1599 Jul;40(7):1950–61. doi:10.1109/TMI.2021.3069663
- 1600 4. Plotka S, Klasa A, Lisowska A, Seliga-Siwecka J, Lipa M, Trzciński T, et al. Deep learning fetal  
1601 ultrasound video model match human observers in biometric measurements. *Phys Med Biol.* 2022  
1602 Feb;67(4):045013. doi:10.1088/1361-6560/ac4d85
- 1603 5. D’Alberty E, Patey O, Smith C, Šalović B, Hernandez-Cruz N, Noble JA, et al. Artificial intelligence-  
1604 enabled prenatal ultrasound for the detection of fetal cardiac abnormalities: a systematic review and  
1605 meta-analysis. *eClinicalMedicine.* 2025 Jun 1;84. doi:10.1016/j.eclinm.2025.103250 PubMed PMID:  
1606 40687738.
- 1607 6. Correia V, Mascarenhas T, Mascarenhas M. Smart Pregnancy: AI-Driven Approaches to Personalised  
1608 Maternal and Foetal Health—A Scoping Review. *J Clin Med.* 2025 Jan;14(19):6974.  
1609 doi:10.3390/jcm14196974
- 1610 7. Abdelwanis M, Alarafati HK, Tammam MMS, Simsekler MCE. Exploring the risks of automation bias  
1611 in healthcare artificial intelligence applications: A Bowtie analysis. *J Saf Sci Resil.* 2024 Dec  
1612 1;5(4):460–9. doi:10.1016/j.jnlssr.2024.06.001
- 1613 8. Giovanola B, Tiribelli S. Beyond bias and discrimination: redefining the AI ethics principle of fairness  
1614 in healthcare machine-learning algorithms. *AI Soc.* 2023 Apr 1;38(2):549–63. doi:10.1007/s00146-022-  
1615 01455-6
- 1616 9. Lendrum TL, Shaffer RK, Heyborne KD. Repeat anatomic surveys performed for an initial incomplete  
1617 study: the sonographer and physician factors. *Am J Obstet Gynecol MFM.* 2022 May 1;4(3).  
1618 doi:10.1016/j.ajogmf.2022.100567 PubMed PMID: 35085846.
- 1619 10. Australasian Society for Ultrasound in Medicine. Guidelines, Policies and Statements: Normal  
1620 Ultrasonic Fetal Measurement Standard [Internet]. 1996. Available from:  
1621 [https://www.asum.com.au/files/public/SoP/curver/Obs-Gynae/Normal-Ultrasonic-Fetal-](https://www.asum.com.au/files/public/SoP/curver/Obs-Gynae/Normal-Ultrasonic-Fetal-Measurements-Standard.pdf)  
1622 [Measurements-Standard.pdf](https://www.asum.com.au/files/public/SoP/curver/Obs-Gynae/Normal-Ultrasonic-Fetal-Measurements-Standard.pdf)
- 1623 11. Alsharid M, Guo X, Men Q, Saha P, Mishra D, Ahuja R, et al. On the public dissemination and open  
1624 sourcing of ultrasound resources, datasets and deep learning models. *Npj Digit Med.* 2025 Nov  
1625 24;8(1):777. doi:10.1038/s41746-025-02162-4
- 1626 12. Fiorentino MC, Moccia S, Cosmo MD, Frontoni E, Giovanola B, Tiribelli S. Uncovering ethical biases  
1627 in publicly available fetal ultrasound datasets. *Npj Digit Med.* 2025 Jun 13;8(1):355.  
1628 doi:10.1038/s41746-025-01739-3

13. Della Ripa S, Santos N, Walker D. AI-enabled obstetric point-of-care ultrasound as an emerging technology in low- and middle-income countries: provider and health system perspectives. *BMC Pregnancy Childbirth*. 2025 Jul 4;25(1):729. doi:10.1186/s12884-025-07796-6
14. Testi M, Fiorentino MC, Ballabio M, Visani G, Ciccozzi M, Frontoni E, et al. FetalMLOps: operationalizing machine learning models for standard fetal ultrasound plane classification. *Med Biol Eng Comput*. 2025 Sep 8. doi:10.1007/s11517-025-03436-5
15. Alabdali SA, Pileggi SF, Cetindamar D. Influential Factors, Enablers, and Barriers to Adopting Smart Technology in Rural Regions: A Literature Review. *Sustainability*. 2023 Jan;15(10):10. doi:10.3390/su15107908
16. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing; 2015. p. 234–41. doi:10.1007/978-3-319-24574-4\_28
17. Amini SM. Head circumference measurement with deep learning approach based on multi-scale ultrasound images. *Multimed Tools Appl*. 2022 Sep 1;81(23):32981–93. doi:10.1007/s11042-022-13107-4
18. Alzubaidi M, Shah U, Agus M, Househ M. FetSAM: Advanced Segmentation Techniques for Fetal Head Biometrics in Ultrasound Imagery. *IEEE Open J Eng Med Biol*. 2024;5:281–95. doi:10.1109/OJEMB.2024.3382487
19. Papageorgiou AT, Walton S, Benson M, Meagher S, Sinkovskaya E, Smith E, et al. EP02.46: An AI system (SonoLyst) achieves expert level performance when categorising images for adherence to ISUOG mid-trimester screening guidelines. *Ultrasound Obstet Gynecol*. 2023;62(S1):117–117. doi:10.1002/uog.26661
20. Salomon L j., Alfirevic Z, Da Silva Costa F, Deter R l., Figueras F, Ghi T, et al. ISUOG Practice Guidelines: ultrasound assessment of fetal biometry and growth. *Ultrasound Obstet Gynecol*. 2019;53(6):715–23. doi:10.1002/uog.20272
21. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Calster BV, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods [Internet]. 2024 Apr 16. doi:10.1136/bmj-2023-078378
22. Sounderajah V, Ashrafian H, Golub RM, Shetty S, Fauw JD, Hooft L, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open*. 2021 Jun 1;11(6):e047709. doi:10.1136/bmjopen-2020-047709 PubMed PMID: 34183345.
23. Vasey B, Novak A, Ather S, Ibrahim M, McCulloch P. DECIDE-AI: a new reporting guideline and its relevance to artificial intelligence studies in radiology. *Clin Radiol*. 2023 Feb 1;Special Issue Section: Artificial Intelligence and Machine Learning78(2):130–6. doi:10.1016/j.crad.2022.09.131
24. Kwong JCC, Khondker A, Lajkosz K, McDermott MBA, Frigola XB, McCradden MD, et al. APPRAISE-AI Tool for Quantitative Evaluation of AI Studies for Clinical Decision Support. *JAMA Netw Open*. 2023 Sep 25;6(9):e2335377. doi:10.1001/jamanetworkopen.2023.35377
25. Alzubaidi M, Agus M, Makhlof M, Anver F, Alyafei K, Househ M. Large-scale annotation dataset for fetal head biometry in ultrasound images. *Data Brief*. 2023 Dec 1;51:109708. doi:10.1016/j.dib.2023.109708
26. van den Heuvel TLA, de Bruijn D, de Korte CL, van Ginneken B. Automated measurement of fetal head circumference using 2D ultrasound images. [Internet]. Zenodo; 2018 [cited 2026 Feb 27]. Available from: <https://hc18.grand-challenge.org/> doi:10.5281/zenodo.1322001
27. Burgos-Artizzu XP, Coronado-Gutiérrez D, Valenzuela-Alcaraz B, Bonet-Carne E, Eixarch E, Crispi F, et al. Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. *Sci Rep*. 2020 Jun 23;10(1):10200. doi:10.1038/s41598-020-67076-5
28. Di Vece C, Mao Z, Avisdris N, Dromey B, Napolitano R, Ben Bashat D, et al. A multicentre benchmark dataset for comprehensive landmark-based fetal ultrasound biometry. *Sci Rep*. 2026 Apr 14;16(1):17405. doi:10.1038/s41598-026-47854-3
29. Niu M, Bai J, Gao Y, Tang Y, Lu Y, Han Z, et al. Maternal-Fetal Ultrasound Video Dataset for End-to-end Intrapartum Biometry and Multi-task Learning. *Sci Data*. 2026 Feb 23;13(1):327. doi:10.1038/s41597-026-06900-5
30. Prabakaran BS, Hamelmann P, Ostrowski E, Shafique M. FPUS23: An Ultrasound Fetus Phantom Dataset With Deep Neural Network Evaluations for Fetus Orientations, Fetal Planes, and Anatomical Features. *IEEE Access*. 2023;11:58308–17. doi:10.1109/ACCESS.2023.3284315

31. Huang Y, Yang X, Liu L, Zhou H, Chang A, Zhou X, et al. Segment anything model for medical images? *Med Image Anal.* 2024 Feb 1;92:103061. doi:10.1016/j.media.2023.103061
32. Zhang Y, Shen Z, Jiao R. Segment anything model for medical image segmentation: Current applications and future directions. *Comput Biol Med.* 2024 Mar 1;171:108238. doi:10.1016/j.compbiomed.2024.108238
33. Sikhakhane K, Rimer S, Gololo M, Ouahada K, Abu-Mahfouz AM. Evaluation of Speckle Noise Reduction Filters and Machine Learning Algorithms for Ultrasound Images. *IEEE Access.* 2024;12:81293–312. doi:10.1109/ACCESS.2024.3411709
34. Bhonsle D, Rizvi T, Mishra S, Sinha GR, Kumar A, Jain VK. Reduction of Ultrasound Images using Combined Bilateral Filter & Median Modified Wiener Filter. In: 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) [Internet]. 2022 [cited 2026 Jul 27]. p. 1–5. Available from: <https://ieeexplore.ieee.org/abstract/document/9807906> doi:10.1109/ICAECT54875.2022.9807906
35. Reza AM. Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement. *J VLSI Signal Process Syst Signal Image Video Technol.* 2004 Aug 1;38(1):35–44. doi:10.1023/B:VLSI.0000028532.53893.82
36. Chen RC, Dewi C, Zhuang YC, Chen JK. Contrast Limited Adaptive Histogram Equalization for Recognizing Road Marking at Night Based on Yolo Models. *IEEE Access.* 2023;11:92926–42. doi:10.1109/ACCESS.2023.3309410
37. Leampar M, Ciocoiu IB. Ultrasound-based Detection of Polycystic Ovary Morphology using YOLO26 on Custom Datasets. In: 2026 International Conference on Development and Application Systems (DAS) [Internet]. 2026 [cited 2026 Jul 28]. p. 235–40. Available from: <https://ieeexplore.ieee.org/abstract/document/11553340> doi:10.1109/DAS69882.2026.11553340
38. Anusree R, Hareendranathan RA, Suhail CPM. Nodule Detection in Breast Ultrasound Using Deep Learning-Based Instance Segmentation. In: 2026 International Conference on Future and Advanced Computing Technologies (ICFACT) [Internet]. 2026 [cited 2026 Jul 28]. p. 1–6. Available from: <https://ieeexplore.ieee.org/abstract/document/11518204> doi:10.1109/ICFACT66887.2026.11518204
39. Li J, Gao Z, Wang C, Pu B, Li K. A rule-guided interpretable lightweight framework for fetal standard ultrasound plane capture and biometric measurement. *Neurocomputing.* 2025 Mar 7;621:129290. doi:10.1016/j.neucom.2024.129290
40. Ou Z, Bai J, Chen Z, Lu Y, Wang H, Long S, et al. RTSeg-net: A lightweight network for real-time segmentation of fetal head and pubic symphysis from intrapartum ultrasound images. *Comput Biol Med.* 2024 Jun 1;175:108501. doi:10.1016/j.compbiomed.2024.108501
41. Fiorentino MC, Villani FP, Di Cosmo M, Frontoni E, Moccia S. A review on deep-learning algorithms for fetal ultrasound-image analysis. *Med Image Anal.* 2023 Jan 1;83:102629. doi:10.1016/j.media.2022.102629
42. Torres HR, Morais P, Oliveira B, Birdir C, Rüdiger M, Fonseca JC, et al. A review of image processing methods for fetal head and brain analysis in ultrasound images. *Comput Methods Programs Biomed.* 2022 Mar 1;215:106629. doi:10.1016/j.cmpb.2022.106629
43. Hekal AA, Amer HM, Moustafa HED, Elnakib A. Automatic measurement of head circumference in fetal ultrasound images using a squeeze atrous pooling UNet. *Biomed Signal Process Control.* 2025 May 1;103:107434. doi:10.1016/j.bspc.2024.107434
44. Zeng Y, Tsui PH, Wu W, Zhou Z, Wu S. Fetal Ultrasound Image Segmentation for Automatic Head Circumference Biometry Using Deeply Supervised Attention-Gated V-Net. *J Digit Imaging.* 2021 Feb 1;34(1):134–48. doi:10.1007/s10278-020-00410-5
45. Oghli MG, Shabanzadeh A, Moradi S, Sirjani N, Gerami R, Ghaderi P, et al. Automatic fetal biometry prediction using a novel deep convolutional network architecture. *Phys Medica Eur J Med Phys.* 2021 Aug 1;88:127–37. doi:10.1016/j.ejmp.2021.06.020 PubMed PMID: 34242884.
46. Maani F, Saeed N, Saleem TJ, Farooq Z, Alasmawi H, Diehl W, et al. FetalCLIP: a visual-language foundation model for fetal ultrasound image analysis. *Npj Digit Med.* 2026 Jun 20. doi:10.1038/s41746-026-02907-9
47. Xian Y, Lampert CH, Schiele B, Akata Z. Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly. *IEEE Trans Pattern Anal Mach Intell.* 2019 Sep;41(9):2251–65. doi:10.1109/TPAMI.2018.2857768
48. Mehrtabar S, Marey A, Desai A, Saad AM, Desai V, Goñi J, et al. Ethical Considerations in Patient Privacy and Data Handling for AI in Cardiovascular Imaging and Radiology. *J Imaging Inform Med.* 2026 Jun 1;39(3):2004–25. doi:10.1007/s10278-025-01656-7

49. Chang K, Balachandar N, Lam C, Yi D, Brown J, Beers A, et al. Distributed deep learning networks among institutions for medical imaging. *J Am Med Inform Assoc.* 2018 Aug 1;25(8):945–54. doi:10.1093/jamia/ocy017
50. Yin X, Zhu Y, Hu J. A Comprehensive Survey of Privacy-preserving Federated Learning: A Taxonomy, Review, and Future Directions. *ACM Comput Surv.* 2021 Jul 13;54(6):131:1-131:36. doi:10.1145/3460427
51. Rehman MH ur, Hugo Lopez Pinaya W, Nachev P, Teo JT, Ourselin S, Cardoso MJ. Federated learning for medical imaging radiology. *Br J Radiol.* 2023 Oct 1;96(1150):20220890. doi:10.1259/bjr.20220890
52. Pillutla K, Kakade SM, Harchaoui Z. Robust Aggregation for Federated Learning. *IEEE Trans Signal Process.* 2022;70:1142–54. doi:10.1109/TSP.2022.3153135
53. Khraisat A, Alazab A, Singh S, Jan T, Jr. Gomez A. Survey on Federated Learning for Intrusion Detection System: Concept, Architectures, Aggregation Strategies, Challenges, and Future Directions. *ACM Comput Surv.* 2024 Oct 7;57(1):7:1-7:38. doi:10.1145/3687124
54. Yadav P, Tam D, Choshen L, Raffel C, Bansal M. TIES-Merging: Resolving Interference When Merging Models [Internet]. arXiv; 2023 [cited 2026 Jul 28]. Available from: <http://arxiv.org/abs/2306.01708> doi:10.48550/arXiv.2306.01708
55. Huang C, Ye P, Chen T, He T, Yue X, Ouyang W. EMR-Merging: Tuning-Free High-Performance Model Merging. In: *Advances in Neural Information Processing Systems* [Internet]. Curran Associates, Inc.; 2024 [cited 2026 Jul 28]. p. 122741–69. Available from: [https://proceedings.neurips.cc/paper\\_files/paper/2024/hash/dda5cac5272a9bcd4bc73d90bc725ef1-Abstract-Conference.html](https://proceedings.neurips.cc/paper_files/paper/2024/hash/dda5cac5272a9bcd4bc73d90bc725ef1-Abstract-Conference.html) doi:10.52202/079017-3900
56. Zeng W, Luo J, Cheng J, Lu Y. Efficient fetal ultrasound image segmentation for automatic head circumference measurement using a lightweight deep convolutional neural network. *Med Phys.* 2022;49(8):5081–92. doi:10.1002/mp.15700
57. Zhao L, Tan G, Pu B, Wu Q, Ren H, Li K. TransFSM: Fetal Anatomy Segmentation and Biometric Measurement in Ultrasound Images Using a Hybrid Transformer. *IEEE J Biomed Health Inform.* 2024 Jan;28(1):285–96. doi:10.1109/JBHI.2023.3328954
58. Zhang J, Petitjean C, Ainouz S. Segmentation-Based vs. Regression-Based Biomarker Estimation: A Case Study of Fetus Head Circumference Assessment from Ultrasound Images. *J Imaging.* 2022 Feb;8(2):23. doi:10.3390/jimaging8020023
59. Herington J, McCradden MD, Creel K, Boellaard R, Jones EC, Jha AK, et al. Ethical Considerations for Artificial Intelligence in Medical Imaging: Data Collection, Development, and Evaluation. *J Nucl Med.* 2023 Dec 1;64(12):1848–54. doi:10.2967/jnumed.123.266080 PubMed PMID: 37827839.
60. Pu B, Li K, Li S, Zhu N. Automatic Fetal Ultrasound Standard Plane Recognition Based on Deep Learning and IIoT. *IEEE Trans Ind Inform.* 2021 Nov;17(11):7771–80. doi:10.1109/TII.2021.3069470
61. Zhu S, Cai J, Xiong R, Zheng L, Ma D. Singular Pooling: A Spectral Pooling Paradigm for Second-Trimester Prenatal Level II Ultrasound Standard Fetal Plane Identification. *IEEE Trans Circuits Syst Video Technol.* 2025 Dec;35(12):12508–23. doi:10.1109/TCSVT.2025.3588395
62. Bashir Z, Lin M, Feragen A, Mikolaj K, Taksøe-Vester C, Christensen AN, et al. Clinical validation of explainable AI for fetal growth scans through multi-level, cross-institutional prospective end-user evaluation. *Sci Rep.* 2025 Jan 15;15(1):2074. doi:10.1038/s41598-025-86536-4
63. Drukker L. Real-time identification of fetal anomalies on ultrasound using artificial intelligence: what's next? *Ultrasound Obstet Gynecol.* 2022;59(3):285–7. doi:10.1002/uog.24869
64. Ziller A, Mueller TT, Stieger S, Feiner LF, Brandt J, Braren R, et al. Reconciling privacy and accuracy in AI for medical imaging. *Nat Mach Intell.* 2024 Jul;6(7):764–74. doi:10.1038/s42256-024-00858-y
65. Elmas G, Dar SUH, Korkmaz Y, Ceyani E, Susam B, Ozbey M, et al. Federated Learning of Generative Image Priors for MRI Reconstruction. *IEEE Trans Med Imaging.* 2023 Jul;42(7):1996–2009. doi:10.1109/TMI.2022.3220757
66. Liu Z, Alavi A, Li M, Zhang X. Self-Supervised Contrastive Learning for Medical Time Series: A Systematic Review. *Sensors.* 2023 Apr 23;23(9):4221. doi:10.3390/s23094221 PubMed PMID: 37177423; PubMed Central PMCID: PMC10181273.
67. Schott B, Pinchuk D, Santoro-Fernandes V, Klaneček Ž, Rivetti L, Deatsch A, et al. Uncertainty quantification via localized gradients for deep learning-based medical image assessments. *Phys Med Biol.* 2024 Jul;69(15):155015. doi:10.1088/1361-6560/ad611d
68. Cui K, Liu W, Wang D. Interpretable diagnosis of breast lesions in ultrasound imaging using deep multi-stage reasoning. *Phys Med Biol.* 2024 Oct;69(21):215025. doi:10.1088/1361-6560/ad869f

69. Zhao H, Zheng Q, Teng C, Yasrab R, Drukker L, Papageorgiou AT, et al. Memory-based unsupervised video clinical quality assessment with multi-modality data in fetal ultrasound. *Med Image Anal.* 2023 Dec 1;90:102977. doi:10.1016/j.media.2023.102977
70. Li X, Zhang H, Yue J, Yin L, Li W, Ding G, et al. A multi-task deep learning approach for real-time view classification and quality assessment of echocardiographic images. *Sci Rep.* 2024 Sep 3;14(1):20484. doi:10.1038/s41598-024-71530-z
71. Albahri AS, Duham AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Inf Fusion.* 2023 Aug 1;96:156–91. doi:10.1016/j.inffus.2023.03.008
72. Raghavan K, B S, v K. Attention guided grad-CAM: an improved explainable artificial intelligence model for infrared breast cancer detection. *Multimed Tools Appl.* 2024 Jun 1;83(19):57551–78. doi:10.1007/s11042-023-17776-7
73. Cho HC, Sun S, Park SW, Kwon JY, Seo JK. Artificial Intelligence for Fetal Ultrasound. In: Seo JK, editor. *Deep Learning and Medical Applications* [Internet]. Singapore: Springer Nature; 2023 [cited 2026 Jul 28]. p. 215–81. Available from: [https://doi.org/10.1007/978-981-99-1839-3\\_5](https://doi.org/10.1007/978-981-99-1839-3_5) doi:10.1007/978-981-99-1839-3\_5
74. Fei N, Lu Z, Gao Y, Yang G, Huo Y, Wen J, et al. Towards artificial general intelligence via a multimodal foundation model. *Nat Commun.* 2022 Jun 2;13(1):3094. doi:10.1038/s41467-022-30761-2
75. Wang B, Wang Y, Chu Y, Zhang K, Liu L, Zhang K, et al. AbVLM-Q: intelligent quality assessment for abdominal ultrasound standard planes via vision-language modeling. *BMC Med Imaging.* 2025 Aug 23;25(1):344. doi:10.1186/s12880-025-01885-w
76. Li Z, Li M, Wang W, Huang Q. Ultrasound report generation with fuzzy knowledge and multi-modal large language model. *Expert Syst Appl.* 2025 Nov 1;292:128555. doi:10.1016/j.eswa.2025.128555
77. Di Vece C, Mao Z, Avisdris N, Dromey B, Napolitano R, Ben Bashat D, et al. A multicentre benchmark dataset for comprehensive landmark-based fetal ultrasound biometry [Internet]. University College London; 2025 [cited 2026 Jul 15]. p. 1588458514 Bytes. Available from: [https://rdr.ucl.ac.uk/articles/dataset/A\\_multi-centre\\_multi-device\\_benchmark\\_dataset\\_for\\_landmark-based\\_comprehensive\\_fetal\\_biometry/30819911](https://rdr.ucl.ac.uk/articles/dataset/A_multi-centre_multi-device_benchmark_dataset_for_landmark-based_comprehensive_fetal_biometry/30819911) doi:10.5522/04/30819911

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AdditionalFile1.pdf](#)
- [AdditionalFile2.pdf](#)
- [AdditionalFile3.pdf](#)
- [AdditionalFile4.pdf](#)