Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Image-only ultrasound AI for dating-independent delivery-date prediction externally validated in four Chilean hospitals

Rodrigo Alliende-Ferrada

rodalliende@gmail.com

Hospital Padre Hurtado

Ignacia Inostroza

Hospital Barros Luco Trudeau

Jocelyn Stern

Hospital y CRS El Pino

Cristobal Pérez-Cotapos

Pontificia Universidad Católica de Chile

Andrés Ruz

Hospital Padre Hurtado

Article

Keywords:

Posted Date: June 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9999161/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Accurate delivery-date estimation underpins obstetric care, yet conventional dating performs worst where it is most needed: late-presenting pregnancies without reliable early ultrasound or menstrual history, common in public-sector and migrant populations. We report an independent, multicentre external validation of a prospectively deployed, image-only ultrasound artificial-intelligence (AI) system that predicts days to delivery from standard ultrasound frames alone—without gestational age, biometry, demographics, annotations, or any dating input. Across 3,788 singleton studies from 2,550 pregnancies at four Chilean public hospitals, the AI predicted the delivery date with a mean absolute error (MAE) of 6.9 days (95% confidence interval [CI] 6.7–7.0). Critically, its accuracy was undiminished where conventional dating is structurally weakest: in pregnancies lacking reliable first-trimester dating the AI achieved an MAE of 6.4 days, whereas estimation from the established due date deteriorated to 13.2 days. The model's high-coverage prediction intervals were well calibrated (96.5% empirical coverage of the approximately 94% interval), supporting a calibrated delivery-date range rather than a single date. Head-to-head within the same studies, the image-only estimate was more accurate than the due date as used in practice in 69% of pregnancies; prespecified sensitivity analyses showed that part of this margin reflects correction of a systematic, population-specific bias in the 280-day convention, with a smaller residual advantage in dispersion. This dating-independent, calibrated estimate could narrow a persistent equity gap for the patients whom conventional dating serves least well.

Introduction

The timing of birth is one of the most consequential quantities in obstetrics. The estimated date of delivery (EDD) anchors gestational-age assignment and the interpretation of fetal growth, the timing of antenatal corticosteroids and fetal surveillance, and decisions about induction, referral, and neonatal-intensive-care planning. Errors in dating propagate through pregnancy management and carry their greatest consequence near the boundaries that define preterm, early-term, and post-term birth.[1]

Conventional dating relies on the last menstrual period, first-trimester crown–rump length (CRL), assisted-reproduction dates, or later biometry when early dating is lacking.[2,3] It performs well when reliable early dating exists, but a large share of patients in public systems present late, have an uncertain menstrual history, or lack a first-trimester scan—and a pregnancy without ultrasound confirmation before 22 0/7 weeks is regarded as suboptimally dated. Moreover, the 280-day convention itself overestimates gestational length in many populations and varies with ethnicity and parity[4,5], so even an assigned due date is a biased predictor of when birth will occur.[6,7] A clinical—and equity—gap therefore persists precisely for the patients in whom an accurate, dating-independent estimate of delivery timing would be most valuable.

Deep learning offers a different route. Rather than depending on a few manually measured biometric variables, an image-only model can learn directly from pixel-level patterns across the entire ultrasound frame, without gestational age, biometry, or operator annotations as inputs, and can be trained against

the calendar date of birth itself. The Perinatal Artificial Intelligence in Ultrasound (PAIR) study established the feasibility of predicting days to delivery from images alone at a single United States centre and explicitly designated calibration of the predicted distribution for future work.[8] Three questions follow naturally for translation: whether the approach transports to an independent health system and to heterogeneous real-world imaging; whether its predicted distribution is well calibrated; and—most importantly for clinical adoption—whether it improves on the existing standard of care, and where.

Here we address those questions. We conducted an independent, multicentre external validation of a prospectively deployed image-only ultrasound AI system across four public hospitals in Chile, a health system and geography distinct from where the model was developed, in an evaluation performed independently of the developer. We (i) quantified delivery-date prediction accuracy; (ii) compared the AI head-to-head, within the same studies, against estimation based on the established due date, with prespecified sensitivity analyses to interrogate the fairness of that comparison; (iii) assessed the calibration of the model's prediction intervals; and (iv) examined performance in the subgroup of pregnancies that lack reliable early dating—where conventional methods are weakest and an image-only estimate could add the most clinical value—and across sites, gestational-age windows, and comorbidities.

Results

Study population

The analytic cohort comprised 3,788 singleton ultrasound studies from 2,550 pregnancies in 2,547 women, performed between March 2021 and November 2025 at four Chilean public hospitals—Barros Luco Trudeau, Padre Hurtado, El Pino, and Dr. Luis Tisné Brousse (Table 1, Fig. 1). The population reflected the real-world, higher-risk case mix of public maternity care: roughly one-third of women were migrants (878/2,550; 34.4%), most commonly from Venezuela (14.6%) and Haiti (8.4%), and pregnancies frequently carried clinically relevant comorbidity, including suspected fetal growth restriction (FGR) (16.6%), gestational diabetes (12.8%), and prior caesarean (13.9%). Each study contained a median of 14 ultrasound frames (interquartile range [IQR] 10–22), acquired on heterogeneous equipment across sites, and 852 pregnancies (33.4%) contributed more than one study, motivating the pregnancy-clustered and one-scan-per-pregnancy analyses below. Scans spanned a wide gestational-age range (mean 30.2 weeks; 34% before 28 weeks), and 95.8% were cephalic at imaging. Overall, 188 of 2,550 pregnancies ended in preterm birth (7.4%; corresponding to 262 of 3,788 studies, 6.9%).

Table 1

Characteristics of the study population, overall and by dating certainty. Values are mean ± standard deviation, median [IQR], or n (%).

Characteristic	Overall	Reliable CRL dating	Late/uncertain dating
Pregnancy- and patient-level	*n = 2,550*	*n = 2,231*	*n = 319*
Maternal age, years	30.4 ± 6.4	30.4 ± 6.5	30.3 ± 6.3
Advanced maternal age ≥ 35 y, n (%)	723 (28.4)	643 (28.8)	80 (25.1)
Adolescent pregnancy, n (%)	106 (4.2)	91 (4.1)	15 (4.7)
Parity, median [IQR]	1 [0–2]	1 [0–2]	1 [0–2]
Chilean nationality, n (%)	1,672 (65.6)	1,484 (66.5)	188 (58.9)
Migrant, n (%)	878 (34.4)	747 (33.5)	131 (41.1)
– Venezuelan, n (%)	372 (14.6)	322 (14.4)	50 (15.7)
– Haitian, n (%)	213 (8.4)	165 (7.4)	48 (15.0)
Maternal BMI, kg/m² (a)	34.4 ± 7.3	34.4 ± 7.3	34.5 ± 7.0
Gestational diabetes, n (%)	326 (12.8)	289 (13.0)	37 (11.6)
Suspected fetal growth restriction, n (%) (b)	424 (16.6)	363 (16.3)	61 (19.1)
Prior caesarean, n (%)	354 (13.9)	304 (13.6)	50 (15.7)
Chronic hypertension, n (%)	211 (8.3)	182 (8.2)	29 (9.1)
Preeclampsia, n (%)	182 (7.1)	156 (7.0)	26 (8.2)
Gestational age at birth, weeks	38.5 ± 1.6	38.6 ± 1.6	38.3 ± 1.6
Preterm birth < 37 weeks, n (%)	188 (7.4)	140 (6.3)	48 (15.0)
Spontaneous onset of labour, n (%)	1,442 (56.5)	1,241 (55.6)	201 (63.0)
Caesarean delivery (any), n (%) (c)	1,131 (44.4)	995 (44.6)	136 (42.6)
Birth weight, g (d)	3,248 ± 578	3,263 ± 580	3,145 ± 556
Perinatal death, n (%)	27 (1.1)	24 (1.1)	3 (0.9)
Study-level	*n = 3,788*	*n = 3,332*	*n = 456*
Gestational age at scan, weeks	30.2 ± 6.8	30.1 ± 7.0	30.9 ± 5.6
Images per study, median [IQR]	14 [10–22]	14 [10–23]	15 [10–21]

Characteristic	Overall	Reliable CRL dating	Late/uncertain dating
Cephalic presentation, n (%)	3,629 (95.8)	3,197 (95.9)	432 (94.7)

Studies were performed March 2021–November 2025. Site distribution (studies): Barros Luco Trudeau 1,563 (41.3%), Padre Hurtado 1,310 (34.6%), El Pino 527 (13.9%), Dr. Luis Tisné Brousse 388 (10.2%). The 2,550 pregnancies arose from 2,547 women (3 contributed 2 pregnancies); 852 (33.4%) contributed more than one study. No statistical tests compare strata, in line with reporting guidance for descriptive baseline tables. BMI, body-mass index; CRL, crown–rump length; IQR, interquartile range. (a) Derived from recorded weight and height of non-standardised timing (available for 2,523 pregnancies); interpret cautiously. (b) Phenotypes derived from coded clinical diagnoses. (c) Mode of delivery undetermined for 154 pregnancies (6.0%); percentages use the full denominator. (d) Available for 2,310 pregnancies.

Reliable early dating by first-trimester CRL was available for 2,231 of 2,550 pregnancies (87.5%; 3,332 studies), whereas 319 pregnancies (12.5%; 456 studies) were dated by later biometry, reflecting entry to care without a first-trimester ultrasound. The two strata differed in clinically meaningful ways: the late/uncertain-dating group contained proportionally more migrant women (41.1% vs 33.5%, driven largely by a doubling of Haitian nationality, 15.0% vs 7.4%) and substantially more preterm births (15.0% vs 6.3%), and was scanned slightly later in pregnancy (mean 30.9 vs 30.1 weeks). The patients in whom a dating-independent estimate would be most useful were therefore also those at highest obstetric risk.

Delivery-date accuracy and comparison with conventional dating

Across the 3,788 studies, the AI predicted the actual delivery date with an MAE of 6.9 days (95% CI 6.7–7.0), a median absolute error of 6.0 days, and a root mean squared error (RMSE) of 8.4 days, with a small signed bias of − 1.5 days (actual minus predicted; Table 2, Fig. 2). Estimates fell within 7, 14, and 21 days of delivery in 56.8%, 92.4%, and 98.7% of studies, and catastrophic errors (> 28 days) occurred in only 0.3%.

In a paired, within-study comparison, the AI was substantially more accurate than conventional estimation based on the established due date. Conventional estimation had an MAE of 11.6 days (95% CI 11.2–12.2) and a pronounced systematic bias of − 10.8 days: in this population, birth occurred on average approximately 11 days (about 1.5 weeks) before the 40-week due date. The AI reduced the MAE by 4.8 days (95% CI 4.3–5.3) and, head-to-head on the same images, was the more accurate of the two estimates in 69.2% of studies—roughly seven of every ten pregnancies (two-sided Wilcoxon signed-rank test, $P < 0.001$; Table 2). It increased the proportion of estimates within 7 days by about half (56.8% vs 37.6%) and reduced catastrophic errors (> 28 days) more than tenfold (0.3% vs 4.0%; Supplementary Table S2). Because both estimates are strongly driven by gestational age at the scan, conventional agreement statistics were also high (R^2 0.97 vs 0.89; concordance correlation coefficient 0.98 vs 0.95)

and did not capture this difference, which was concentrated in bias, within-window accuracy, and the catastrophic-error rate.

Table 2

Image-only AI versus conventional due-date estimation of delivery date: head-to-head accuracy, overall and by subgroup. Each row is a paired, within-study comparison in the indicated population. Point prediction = model median (quantile 0.50); conventional estimate = the established due date projected to the scan (280 days minus gestational age at scan). Full secondary metrics (median absolute error, RMSE, R^2, concordance, within-7-day and within-21-day accuracy, and the > 28-day catastrophic-error rate) are given in Supplementary Table S2.

Population (n)	AI MAE, d (95% CI)	Conventional MAE, d (95% CI)	ΔMAE, d (95% CI)	AI more accurate, %	Within 14 d, % (AI / conv.)	Bias, d (AI / conv.)
All singletons (3,788)	6.9 (6.7–7.0)	11.6 (11.2–12.2)	−4.8 (−5.3, −4.3)	69.2	92.4 / 72.9	−1.5 / −10.8
Reliable CRL dating (3,332)	6.9 (6.7–7.1)	11.4 (10.9–12.0)	−4.5 (−5.0, −4.0)	68.6	92.9 / 74.0	−1.5 / −10.6
Late/uncertain dating (456)	6.4 (5.9–7.0)	13.2 (12.1–14.3)	−6.7 (−8.0, −5.6)	73.2	88.4 / 64.3	−1.2 / −12.3
Term, ≥ 37 wk (3,526)	6.6 (6.4–6.7)	9.7 (9.4–10.0)	−3.2 (−3.5, −2.9)	66.9	93.9 / 78.3	−1.1 / −8.9
Preterm, < 37 wk (262)	10.8 (9.6–12.0)	37.5 (33.9–41.3)	−26.6 (−29.6, −23.9)	99.6	71.4 / 0.0	−7.4 / −37.5
Earliest scan per pregnancy (2,550)	6.9 (6.7–7.1)	11.2 (10.8–11.6)	−4.3 (−4.7, −3.9)	67.0	92.3 / 74.9	−1.2 / −10.3
Latest scan per pregnancy (2,550)	7.0 (6.8–7.2)	11.2 (10.8–11.6)	−4.2 (−4.6, −3.9)	66.5	92.2 / 74.9	−1.6 / −10.3

ΔMAE is the mean within-study difference in absolute error (AI minus conventional); negative values favour the AI, and all ΔMAE comparisons were significant at $P < 0.001$ (two-sided Wilcoxon signed-rank test on per-study absolute errors). “AI more accurate, %” is the percentage of studies in which the AI’s absolute error was smaller than the conventional estimate’s—the head-to-head win rate. “Within 14 d” is the percentage of studies whose estimate fell within 14 days of the actual delivery date. Bias is the mean signed error (actual minus predicted); negative values indicate that birth occurred earlier than predicted, and the conventional estimate’s large negative bias reflects this population’s gestational length being about 1.5 weeks shorter than the 40-week assumption. All 95% confidence intervals are from a non-parametric bootstrap resampling pregnancies (3,000 resamples). The conventional estimate cannot

anticipate preterm delivery by construction, which explains its error in the preterm stratum. CI, confidence interval; CRL, crown–rump length; MAE, mean absolute error.

Performance where early dating is unreliable

The AI's advantage was greatest where conventional dating is structurally weakest (Table 2, Fig. 3). In the 456 studies dated by later biometry, the AI achieved an MAE of 6.4 days (95% CI 5.9–7.0)—no worse, and numerically better, than the 6.9 days in CRL-dated studies—whereas conventional estimation deteriorated to an MAE of 13.2 days (95% CI 12.1–14.3). The paired advantage widened accordingly: the AI was the more accurate estimate in 73.2% of uncertainly dated studies—nearly three of every four—reducing the MAE by 6.7 days (95% CI 5.6–8.0). Because uncertainly dated pregnancies were scanned slightly later, we compared the strata within matched gestational-age-at-scan windows; AI error remained comparable (14–<28 weeks, 7.3 vs 7.1 days; ≥28 weeks, 6.0 vs 6.8 days; uncertain vs reliable dating), confirming that the model's accuracy did not depend on the availability of reliable first-trimester dating. This robustness is mechanistically coherent: the model is trained against the calendar date of birth, which is independent of how a pregnancy was dated clinically, and therefore does not inherit the error of an uncertain menstrual history or a late dating scan.

The contrast was most extreme for preterm births, where the established due date is uninformative by construction (it assumes a 40-week delivery): conventional estimation predicted delivery 37.5 days too late on average, with no estimate within 21 days, whereas the AI retained an MAE of 10.8 days and was more accurate in 99.6% of cases (Table 2). As expected, AI accuracy was higher for term than preterm births (MAE 6.6 vs 10.8 days), consistent with the contribution of acute obstetric events to preterm timing.

Robustness of the comparison

Because the conventional estimate carried a large systematic bias, we performed two prespecified sensitivity analyses to interrogate the head-to-head result (Supplementary Table S1). First, restricting the comparison to spontaneous-onset deliveries—the convention in studies of due-date accuracy, which removes the influence of clinically scheduled delivery—left the AI's advantage essentially unchanged (MAE 6.3 vs 11.6 days; smaller error in 69.1%), and notably the conventional bias was identical in spontaneous and iatrogenic deliveries (− 10.8 vs − 10.9 days), indicating that the due date's over-estimation reflects a genuinely shorter gestational length in this population rather than scheduled early delivery. Within spontaneous deliveries, the AI's advantage was again largest where dating was uncertain (MAE 5.6 vs 14.3 days). Second, we compared the AI against an "oracle" bias-corrected due date, recentred using the cohort's own mean gestational length—a best case for the conventional approach that uses information unavailable prospectively. Against this oracle comparator the AI's overall advantage attenuated to 0.7 days and was no longer significant ($P = 0.19$), but a modest advantage persisted in spontaneous-onset deliveries (MAE 6.3 vs 8.2 days; $P < 0.001$). Thus a large part of the AI's

superiority over the due date as used in practice reflects correction of a systematic, population-specific bias, augmented by a smaller advantage in dispersion; critically, the AI achieves this while requiring no dating input at all and while providing a calibrated range, neither of which the conventional approach can offer.

Calibration of prediction intervals

The model's predicted quantile intervals were well calibrated at high coverage (Table 3, Fig. 4). The approximately 94% interval (quantiles 0.03–0.97) covered 96.5% of actual delivery dates (95% CI 96.0–97.1; absolute calibration error 2.5 percentage points; mean width 41.4 days), and the 90% interval (quantiles 0.05–0.95) covered 89.8% (calibration error 0.2 percentage points; mean width 30.7 days). The narrower central intervals departed modestly from nominal (80% interval, 77.0%; 60% interval, 65.3%), with high-coverage intervals closest to nominal. Calibration was preserved across dating strata: the approximately 94% interval covered 96.8% of CRL-dated and 94.3% of biometry-dated deliveries. These results support reporting the model output as a calibrated delivery-date range rather than a single predicted date—particularly valuable when a single, dating-anchored due date is unavailable or unreliable.

Table 3
Calibration of the model's predicted delivery-date intervals (all singletons, n = 3,788).

Interval (quantiles)	Nominal, %	Empirical coverage, % (95% CI)	Mean width, days	Abs. calibration error, pp
Central 60% (0.20–0.80)	60	65.3 (63.8–66.8)	17.7	5.3
Central 80% (0.10–0.90)	80	77.0 (75.7–78.4)	22.5	3.0
Central 90% (0.05–0.95)	90	89.8 (88.8–90.7)	30.7	0.2
Approximately 94% (0.03–0.97)	94	96.5 (96.0–97.1)	41.4	2.5

The 95% CI is a binomial confidence interval; pp, percentage points; absolute calibration error = |empirical − nominal|. Calibration by dating certainty: the approximately 94% interval covered 96.8% (reliable CRL) and 94.3% (biometry) of deliveries, and the 90% interval covered 90.1% and 87.1%, respectively.

Transportability and robustness across sites, gestation, and comorbidity

Performance was consistent across the four hospitals despite heterogeneous equipment and acquisition, with site-level MAE ranging from 6.0 to 7.7 days and high-coverage interval calibration

preserved at every site (96–99% coverage of the approximately 94% interval; Fig. 5). Accuracy was stable across gestational-age-at-scan windows (MAE 7.7, 7.0, 6.6, and 6.7 days for scans at < 20, 20–<28, 28–<34, and ≥ 34 weeks) and across major comorbidity subgroups (MAE 6.8–8.1 days), including measured obesity (n = 2,715; MAE 6.8 days), a condition that degrades ultrasound image quality; the AI outperformed conventional estimation in every comorbidity subgroup examined. Results were robust to repeated scans within a pregnancy: restricting to a single study per pregnancy yielded an MAE of 6.9 days (earliest scan) and 7.0 days (latest scan), with the head-to-head advantage preserved (Table 2, Fig. 5).

Discussion

In this independent, multicentre, real-world evaluation across four Chilean public hospitals, an image-only ultrasound AI system predicted the actual delivery date with a mean absolute error of 6.9 days and placed more than 90% of deliveries within 14 days of birth, using standard-of-care frames alone—without gestational age, biometry, demographics, or sonographer annotations. Three findings move this beyond a confirmatory external validation. First, in a paired comparison within the same studies, the AI was more accurate than the established due date as it is actually used at the bedside—in roughly seven of every ten pregnancies—an advantage that, as we decompose below, combines correction of a population-specific dating bias with a genuine, if smaller, gain in precision. Second, its advantage was largest precisely where conventional dating is weakest—in pregnancies without reliable early dating—and it performed well on later scans and in a high-risk, migrant, imaging-challenging population. Third, its high-coverage prediction intervals were well calibrated, allowing the output to be reported as a credible range, which directly addresses the question PAIR left open.

The head-to-head win rate deserves emphasis, because it is a per-pregnancy rather than an aggregate measure. Population-level error summaries can obscure how often one method actually beats another in an individual patient; here, presented with both estimates for the same pregnancy, the image-only prediction was the more accurate one to act on in roughly seven of every ten cases, and in nearly three of four when early dating was unreliable. For the real bedside decision—whether to trust an uncertain due date or an image-only estimate for a given woman—this is the directly relevant quantity, and it favoured the AI most precisely where confidence in conventional dating is already lowest.

The central clinical implication is that an image-only estimate is independent of dating. Because the model is trained against the calendar date of birth and learns from pixel-level image patterns, it does not inherit the error of an uncertain last menstrual period or a late dating scan; accordingly, its accuracy was undiminished in biometry-dated pregnancies and within matched gestational-age windows. The decisions that hinge on delivery timing—when to administer antenatal corticosteroids, how intensively to surveil, whether an apparently post-term pregnancy genuinely warrants induction, and how to plan delivery or referral—are exactly those that go wrong when the underlying dating is wrong, and dating errors concentrate in patients who present late, have no first-trimester scan, or carry an uncertain menstrual history. In our cohort these were also the highest-risk pregnancies and were enriched for

migrant and, in particular, Haitian women. A reliable, calibrated, dating-free estimate of delivery timing would therefore most change management precisely where conventional dating fails, making this an equity-relevant capability and not merely a technical one.

How such a tool might fit clinical workflows follows from these properties, albeit as a cautious, hypothesis-generating suggestion. Requiring only standard-of-care frames, it could serve as a second read at a routine scan—most useful where first-trimester dating is absent, late, or uncertain, which in this cohort coincided with migrant and otherwise vulnerable patients. Such use would be an adjunct to established dating, not a replacement, and would need prospective evaluation against clinical outcomes.

We were deliberately rigorous about the comparison. The established due date over-estimated time to delivery by approximately 11 days, and—because this bias was essentially identical in spontaneous and scheduled deliveries (− 10.8 vs − 10.9 days)—it reflects a genuinely shorter gestational length in this population rather than an artefact of iatrogenic delivery, consistent with evidence that the 280-day convention overestimates gestation in many non-European populations. The AI's raw advantage therefore decomposes into two parts: a large component that corrects this systematic, population-specific bias, and a smaller component that reflects tighter dispersion. When we removed the bias by recentring the due date on the cohort's own mean signed error—an "oracle" that uses information no clinician has prospectively—the overall difference shrank to 0.7 days and lost statistical significance (P = 0.19), while a modest dispersion advantage persisted in spontaneous-onset deliveries (1.9 days; P < 0.001) and the AI alone supplied a calibrated interval. We regard this decomposition as a strength rather than a caveat: such recentring is itself only possible once a population's mean gestation is already known, whereas the AI requires no dating, no population-specific calibration, and no menstrual history. Its practical value does not rest on out-performing an optimally recalibrated due date; it rests on delivering accurate, calibrated predictions for the very patients for whom a trustworthy due date does not exist.

Read alongside PAIR, these findings trace a coherent translational arc. PAIR established, at a single United States centre, that days to delivery can be predicted from ultrasound images alone; the present study shows that the same image-only approach retains its accuracy in an independent, multi-institutional public-hospital system and on heterogeneous equipment, adds the calibration layer PAIR set aside, quantifies superiority over the standard of care, and demonstrates preserved performance where dating is least certain. Independent external validation of this kind strengthens, rather than competes with, the original observation and moves the technology closer to clinical use.

This study has several limitations, each pointing to a clear next step. The AI system is proprietary, so we could not inspect its architecture or internal representations or provide model-level explainability; accordingly, we frame this work as a pragmatic, real-world evaluation of a deployed system rather than the description of a new model. The choice of comparator likewise warrants a precise statement. The deployment dataset did not retain the stored fetal biometric measurements (biparietal diameter, head and abdominal circumference, femur length) needed to recompute a contemporaneous Hadlock

gestational age from the same frames, so we benchmarked the AI against the operative clinical estimate rather than against same-scan ultrasound re-dating. This was a deliberate and clinically meaningful choice: the established due date, projected forward to the scan, is precisely the estimate clinicians act on at the bedside and the baseline an image-only tool would have to improve upon in practice. It is, however, not equivalent to same-scan biometric dating, and our design therefore cannot separate the model's contribution from that of contemporaneous biometry measured on the same image. The single most important next step is therefore a prospective, protocol-driven head-to-head in which fetal biometry is captured and adjudicated at the same scan, allowing the image-only prediction and a contemporaneous Hadlock-based estimate to be compared directly on identical frames; until such data exist, benchmarking against the operative due date remains the comparison that reflects real bedside decision-making. The preterm contrast should likewise be read in light of the fact that a 40-week due date cannot, by construction, anticipate preterm delivery.

The study was retrospective, a post-hoc analysis of a prospectively deployed system; analyses were conducted at the study level, with a pregnancy-clustered bootstrap and one-scan-per-pregnancy sensitivity analyses used to account for repeated scans. The primary cohort comprised the eligible singleton studies with the complete data required for the dating-certainty classification and the head-to-head comparison (Fig. 1); studies excluded for missing these fields could differ from those analysed, although the primary accuracy estimate was stable across the earliest- and latest-scan sensitivity populations. Maternal body-mass index was derived from recorded weight and height of non-standardised timing and should be interpreted cautiously. Finally, the cohort was drawn from public hospitals in a single country, so generalizability to private systems, higher-income referral networks, and other ultrasound workflows will require further validation.

Looking ahead, the natural next steps are a prospective evaluation in routine care, an audited same-scan comparison against biometry-based dating, multi-country validation, and explainability analyses contingent on model access, together with studies of how a calibrated, dating-independent estimate changes clinical decisions and outcomes. By extending an established image-only approach into independent, multi-site deployment, quantifying its advantage over the standard of care, and adding calibrated uncertainty, this study supports the continued development of ultrasound AI as a practical adjunct for delivery-date estimation—particularly for the patients whom conventional dating serves least well.

Methods

Study design and reporting

This was a multicentre retrospective cohort study, conducted as a post-hoc analysis of a prospectively deployed image-only ultrasound AI system applied to routine obstetric ultrasound at four Chilean public hospitals. The model was not retrained, fine-tuned, or otherwise modified for this analysis, and the data

analysed here were not used in its development or training. Reporting follows the TRIPOD + AI statement for prediction-model studies[10]; the completed checklist is provided as Supplementary Information.

Setting and ethics

The four participating sites—Barros Luco Trudeau, Dr. Luis Tisné Brousse, Padre Hurtado, and El Pino hospitals—are all part of the Chilean public health system and used the same deployed version of the AI system. The study was approved by three independent research ethics committees with jurisdiction over the corresponding hospitals: the Comité Ético Científico of the Servicio de Salud Metropolitano Sur (protocol 123-30102025, approved 11 November 2025; El Pino and Barros Luco Trudeau); the Comité de Ética Científico of the Servicio de Salud Metropolitano Oriente (protocol SSMO 141025-9, approved 14 October 2025; Dr. Luis Tisné Brousse); and the Comité Ético Científico of Padre Hurtado Hospital (internal registry 2025-19, approved 20 August 2025; Resolución Exenta No. 1084, 29 August 2025). All three committees granted a waiver of informed consent given the retrospective design and the use of fully de-identified data, in accordance with the Declaration of Helsinki and Chilean law (Ley 20.120 and Ley 20.584). Patient identifiers were removed from all datasets before analysis.

Data sources and study population

Routine obstetric ultrasound studies performed between 2021 and 2025 were linked to delivery outcomes through a unique study identifier, yielding 4,515 outcome-linked studies from 2,792 pregnancies. Eligibility required a non-missing actual delivery date, a non-missing gestational age at birth, at least three ultrasound images, and a documented dating method for the first ultrasound of the pregnancy. Twin pregnancies (402 studies) were excluded, as were 325 singleton studies (132 pregnancies) lacking data required for the dating-certainty classification and the head-to-head comparison, leaving 3,788 singleton studies from 2,550 pregnancies as the primary analytic population (Fig. 1). To assess robustness to repeated scans, two one-scan-per-pregnancy sensitivity populations were defined, retaining either the earliest or the latest study per pregnancy.

AI system and outcomes

The AI system (Ultrasound AI, Inc., Lexington, KY, USA) accepts standard-of-care ultrasound image frames as input and derives its predictions from pixel-level patterns across the full frame, consistent with the image-only approach introduced in the PAIR study. It does not use gestational age, biometric measurements, demographics, or sonographer annotations as inputs. For each study, the model outputs a predicted distribution of days until delivery, represented by quantiles from 0.03 to 0.97. The primary outcome was the actual delivery date, operationalised as the number of calendar days from the ultrasound study to birth; the point prediction was the model median (quantile 0.50), and the predicted delivery date and its range were obtained by adding the predicted day quantiles to the study date.

Conventional dating comparator

For each study we computed the conventional estimate of days to delivery as the number of days from the scan to the patient's established due date (estimated date of delivery, EDD). The EDD is the date on

which the pregnancy reaches 40 completed weeks under the clinically assigned dating—itself derived, in order of priority, from first-trimester CRL, assisted-reproduction dates, a reliable last menstrual period, or later biometry. Projecting that fixed EDD forward to the scan date is therefore arithmetically equivalent to 280 days minus the gestational age at the scan, expressed in completed days. This is the routine clinical projection of time to delivery once an EDD has been assigned, and the estimate against which an image-only model must demonstrate added value. Prediction error was defined identically for both methods as the actual minus the predicted number of days, so that absolute error, bias, and within-window accuracy are directly comparable. In two prespecified sensitivity analyses we (i) restricted the comparison to spontaneous-onset deliveries—the convention in studies of due-date accuracy, which removes the influence of clinically scheduled delivery—and (ii) compared the AI against a bias-corrected ("oracle") due date recentred by the cohort's mean signed error, explicitly acknowledging that this recentring uses outcome information that would be unavailable prospectively and therefore represents a best case for the conventional approach.

Dating-certainty classification

Each pregnancy was classified by the dating method of its first ultrasound. Pregnancies whose first scan measured first-trimester CRL were considered to have reliable early dating; those first dated by later biometry were considered to have uncertain dating—a pattern typical of late entry to care without a first-trimester scan. Because the uncertain-dating group tended to be scanned later, the strata were additionally compared within matched gestational-age-at-scan windows.

Statistics and reproducibility

Analyses were performed in Python 3.11 (packages pandas, numpy, and scipy). Continuous variables are summarised as mean ± standard deviation (s.d.) or median [IQR]; categorical variables as count (percentage). Unless stated otherwise, the n for each analysis is the number of ultrasound studies; the primary cohort comprised 3,788 studies from 2,550 pregnancies. Prediction accuracy was quantified with mean and median absolute error, RMSE, the coefficient of determination (R^2), Lin's concordance correlation coefficient, signed bias (actual minus predicted), within-window accuracy at 7, 14, and 21 days, and the catastrophic-error rate (absolute error > 28 days).

The image-only AI and the conventional estimate were compared on the same studies (a paired design); the comparison of interest was the per-study difference in absolute error. Because per-study absolute errors were right-skewed and non-normal, as judged by inspection of histograms and normal quantile–quantile plots (Fig. 2c), we used the two-sided Wilcoxon signed-rank test (a non-parametric paired test) rather than a paired t-test; this test generated all reported P values for AI-versus-conventional comparisons. All tests were two-sided with an alpha level of 0.05. For the principal comparison (all singletons, n = 3,788) and for all prespecified subgroup and sensitivity comparisons except the bias-corrected ("oracle") analysis, the Wilcoxon P value was smaller than the smallest value reliably representable and is reported as $P < 0.001$ (for the all-singletons comparison the value was approximately 1×10^{-177}); the bias-corrected comparison was not significant in all singletons ($P = 0.19$)

and was $P < 0.001$ (approximately 2×10^{-13}) in spontaneous-onset deliveries. Subgroup analyses (by dating certainty, gestational age at scan, term versus preterm birth, comorbidity, site, onset of labour, and one-scan-per-pregnancy population) were prespecified and descriptive; because each addresses a distinct, pre-identified clinical question rather than a search across many groupings, and because the principal comparison stands on its own, no multiplicity adjustment was applied and subgroup P values are interpreted as supportive rather than confirmatory.

Confidence intervals for accuracy metrics and for between-method differences were obtained by a non-parametric bootstrap that resampled pregnancies (not individual studies) with replacement (3,000 resamples), thereby accounting for the clustering of multiple studies within a pregnancy; reported 95% CIs are the 2.5th and 97.5th percentiles of the bootstrap distribution, and the same procedure produced the error bars in Figs. 3 and 5. Prediction-interval calibration was assessed by comparing empirical with nominal coverage for the central 60%, 80%, 90%, and approximately 94% intervals, with binomial (Clopper–Pearson) 95% CIs (the error bars in Fig. 4a,b). No statistical power calculation was performed because the study analysed all available outcome-linked studies meeting the eligibility criteria—a fixed, complete sample—rather than a prospectively powered comparison.

Use of large language models

A large language model (Claude; Anthropic) was used to assist with language editing and with the organisation and formatting of the manuscript text. It was not used to generate, analyse, or interpret the study data, and it does not meet authorship criteria and is not an author. All scientific content, analyses, and conclusions are the responsibility of the authors, who reviewed and verified the final manuscript.

Data availability

Individual-level clinical data cannot be made publicly available because they contain potentially identifying information and their sharing is restricted by Chilean data-protection law (Ley 19.628) and by institutional data-use agreements with the participating public hospitals. In accordance with journal policy, the de-identified minimum dataset required to reproduce the findings reported here—comprising, for each study, the model's predicted day quantiles, the conventional due-date estimate, the realised time to delivery and gestational age at birth, and the de-identified covariates used in the subgroup analyses, with all calendar dates and direct identifiers removed—has been deposited under controlled access (Zenodo restricted-access record, accession XXXXXXX [assigned on acceptance]). Access will be granted to qualified researchers for non-commercial scientific use upon submission of a brief analysis plan and execution of a data-use agreement, subject to approval by the participating institutions and the relevant research ethics committees; requests are coordinated by the corresponding author and will receive a decision within eight weeks. This de-identified minimum dataset was also made available to the editors and peer reviewers to enable independent verification of the analyses during review.

Code availability

The AI system evaluated here is a proprietary commercial product (Ultrasound AI, Inc.); its source code, model weights, and training data are not available, and this restriction is stated explicitly in accordance with journal policy on the independent evaluation of third-party tools. Importantly, none of the results reported in this paper depend on access to the model internals: every figure, accuracy metric, calibration estimate, bootstrap confidence interval, and subgroup analysis is reproducible from a de-identified table of per-study predicted day quantiles and delivery outcomes. All analytical code (Python 3.11) implementing these analyses has been deposited under an open-source licence in a public repository (Zenodo, https://doi.org/10.5281/zenodo.XXXXXXX [assigned on acceptance]; mirrored at https://github.com/[org]/[repo]), together with a synthetic example dataset that reproduces the full analysis pipeline end-to-end.

Declarations

Acknowledgements

This study received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The authors thank the clinical, sonography, and medical-records staff of the four participating hospitals—Hospital Barros Luco Trudeau, Hospital Padre Hurtado, Hospital El Pino, and Hospital Dr. Luis Tisné Brousse—whose routine care generated the ultrasound studies and delivery outcomes on which this independent evaluation was based.

Author contributions

R.A.-F. and A.R. conceived and designed the study. C.P.-C. developed the analysis plan and performed the data linkage, statistical analysis, and calibration assessment. I.I. and J.S. provided clinical oversight at the participating hospitals and contributed to data acquisition and curation; R.A.-F. and A.R. also contributed clinical data from their respective hospitals. C.P.-C. and R.A.-F. prepared the figures and tables. A.R. supervised the study and provided senior clinical and methodological guidance. R.A.-F. drafted the manuscript, and all authors critically revised it for important intellectual content. All authors had full access to the study data, take responsibility for the integrity of the data and the accuracy of the analysis, and approved the final version for submission.

Competing interests

The AI system evaluated in this study is a commercial product of Ultrasound AI, Inc. (Lexington, KY, USA). This work was an independent external validation conducted without the developer’s involvement: the company had no role in the study design, data collection, analysis, interpretation of the results, preparation of the manuscript, or the decision to submit it for publication, and it provided no funding, equipment, or material support; access to the system reflected only its routine clinical deployment at the participating hospitals. The authors declare no financial or non-financial competing interests.

References

1. Goldenberg, R. L., Culhane, J. F., Iams, J. D. & Romero, R. Epidemiology and causes of preterm birth. *Lancet* 371, 75–84 (2008). https://doi.org/10.1016/S0140-6736(08)60074-4
2. American College of Obstetricians and Gynecologists, American Institute of Ultrasound in Medicine & Society for Maternal-Fetal Medicine. Methods for estimating the due date. Committee Opinion No. 700. *Obstet. Gynecol.* 129, e150–e154 (2017). https://doi.org/10.1097/AOG.0000000000002046
3. Butt, K. & Lim, K. I. Guideline No. 388 — determination of gestational age by ultrasound. *J. Obstet. Gynaecol. Can.* 41, 1497–1507 (2019). https://doi.org/10.1016/j.jogc.2019.04.010
4. Lawson, G. W. Naegele's rule and the length of pregnancy — a review. *Aust. N. Z. J. Obstet. Gynaecol.* 61, 177–182 (2021). https://doi.org/10.1111/ajo.13253
5. Jukic, A. M., Baird, D. D., Weinberg, C. R., McConnaughey, D. R. & Wilcox, A. J. Length of human pregnancy and contributors to its natural variation. *Hum. Reprod.* 28, 2848–2855 (2013). https://doi.org/10.1093/humrep/det297
6. Nguyen, T. H., Larsen, T., Engholm, G. & Møller, H. Evaluation of ultrasound-estimated date of delivery in 17,450 spontaneous singleton births: do we need to modify Naegele's rule? *Ultrasound Obstet. Gynecol.* 14, 23–28 (1999). https://doi.org/10.1046/j.1469-0705.1999.14010023.x
7. Taipale, P. & Hiilesmaa, V. Predicting delivery date by ultrasound and last menstrual period in early gestation. *Obstet. Gynecol.* 97, 189–194 (2001). https://doi.org/10.1016/S0029-7844(00)01131-5
8. Patel, N. et al. Perinatal artificial intelligence in ultrasound (PAIR) study: predicting delivery timing. *J. Matern. Fetal Neonatal Med.* 38, 2532099 (2025). https://doi.org/10.1080/14767058.2025.2532099
9. Bland, J. M. & Altman, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1, 307–310 (1986). https://doi.org/10.1016/S0140-6736(86)90837-8
10. Collins, G. S. et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385, e078378 (2024). https://doi.org/10.1136/bmj-2023-078378

Figures

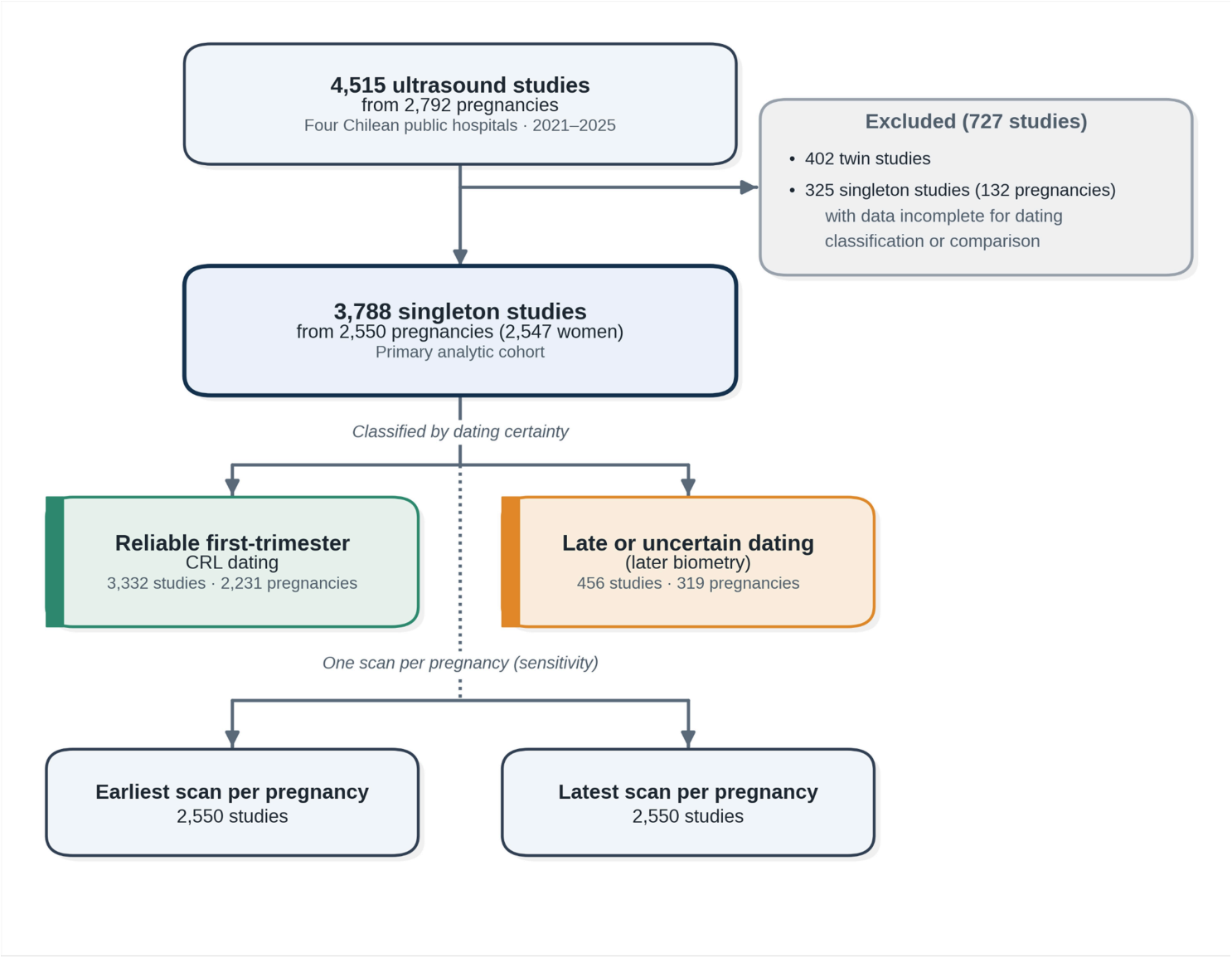


Figure 1

Cohort flow and analytic population. Outcome-linked obstetric ultrasound studies from four Chilean public hospitals (2021–2025) were screened for eligibility; exclusions and counts are shown in each box. The primary analytic cohort comprised 3,788 singleton studies from 2,550 pregnancies (2,547 women), classified by dating certainty—reliable first-trimester crown–rump length dating versus late or uncertain dating by later biometry—with earliest- and latest-scan, one-per-pregnancy populations derived for sensitivity analysis. No error bars are shown (counts).

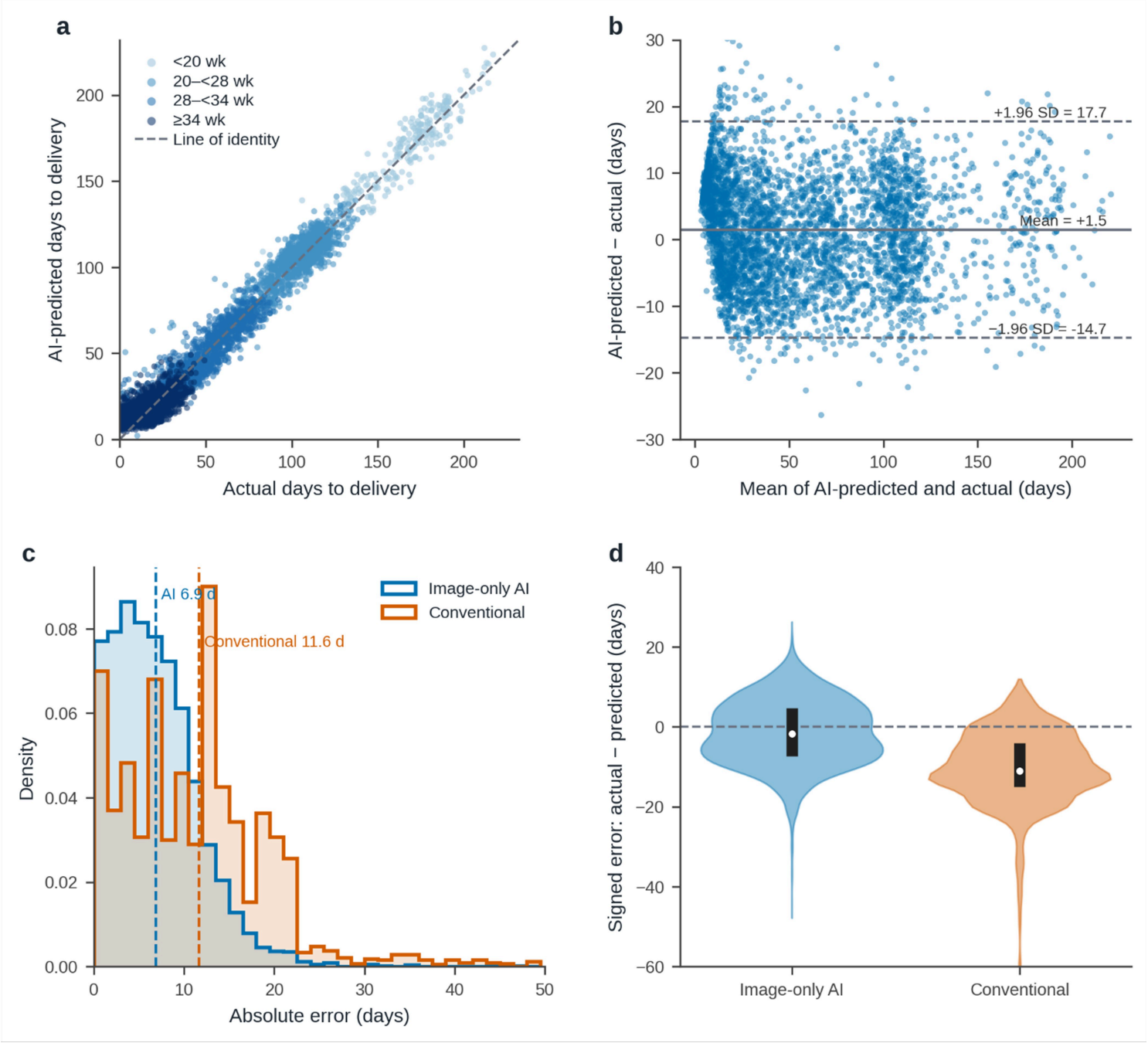


Figure 2

Accuracy of AI delivery-date prediction and comparison with conventional dating. (a) AI-predicted versus actual days from ultrasound to delivery for all 3,788 singleton studies; the dashed line is the line of identity and points are coloured by gestational-age window at the scan (shaded from light to dark blue: <20, 20–<28, 28–<34, and ≥34 weeks). (b) Bland–Altman plot[9] of the difference (AI-predicted minus actual) against the mean of the two; the solid line is the mean difference (1.5 days) and the dashed lines are the 95% limits of agreement (mean ± 1.96 standard deviations). (c) Distributions of per-study absolute error for the image-only AI (blue) and the established due date (orange), with vertical lines marking each mean; the conventional method shows a longer right-hand tail. (d) Signed error (actual minus predicted) by method shown as violins with median and interquartile range; the conventional

estimate has a systematic bias of about −11 days whereas the AI is centred near zero (dashed line at zero).

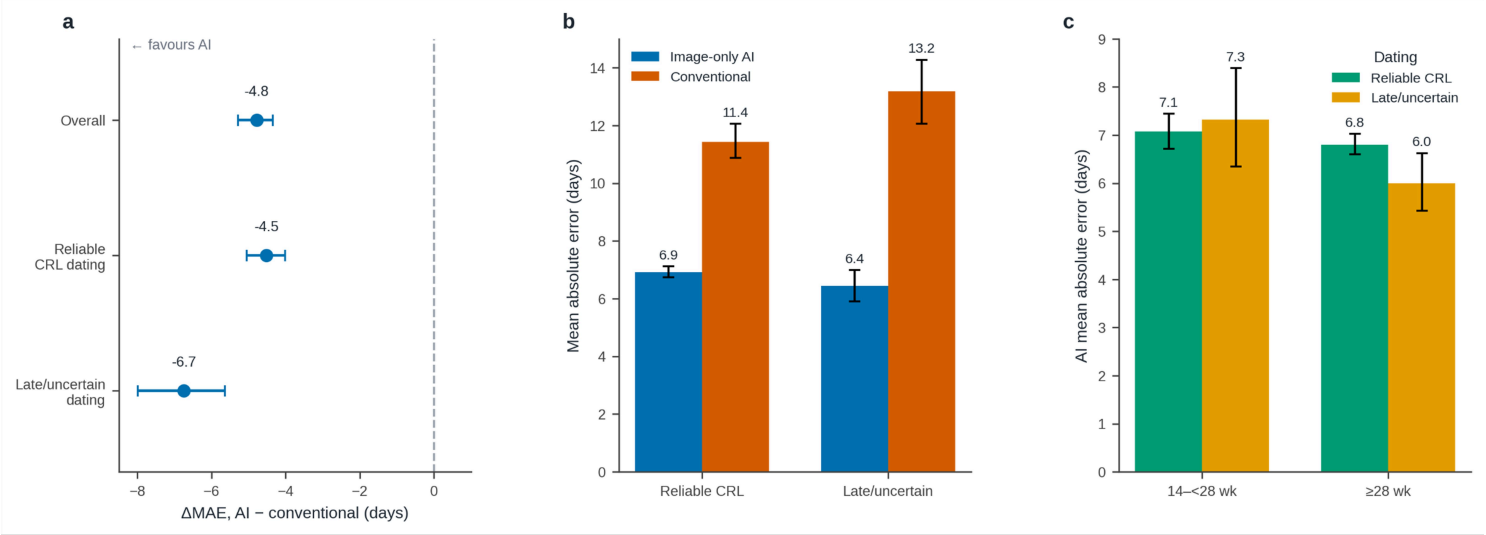


Figure 3

Performance where early dating is unreliable. (a) Difference in mean absolute error between the image-only AI and conventional estimation (blue circles), overall and by dating certainty; values left of the dashed zero line favour the AI, and the advantage widens for late/uncertain dating. (b) Mean absolute error for the image-only AI (blue) and conventional estimation (orange) in reliably and uncertainly dated studies; the AI is flat across strata while the conventional error increases. (c) AI mean absolute error by dating certainty (green, reliable; amber, late/uncertain) within matched gestational-age-at-scan windows. Error bars in all panels are 95% confidence intervals from a non-parametric bootstrap that resampled pregnancies (not individual studies) with replacement (3,000 resamples).

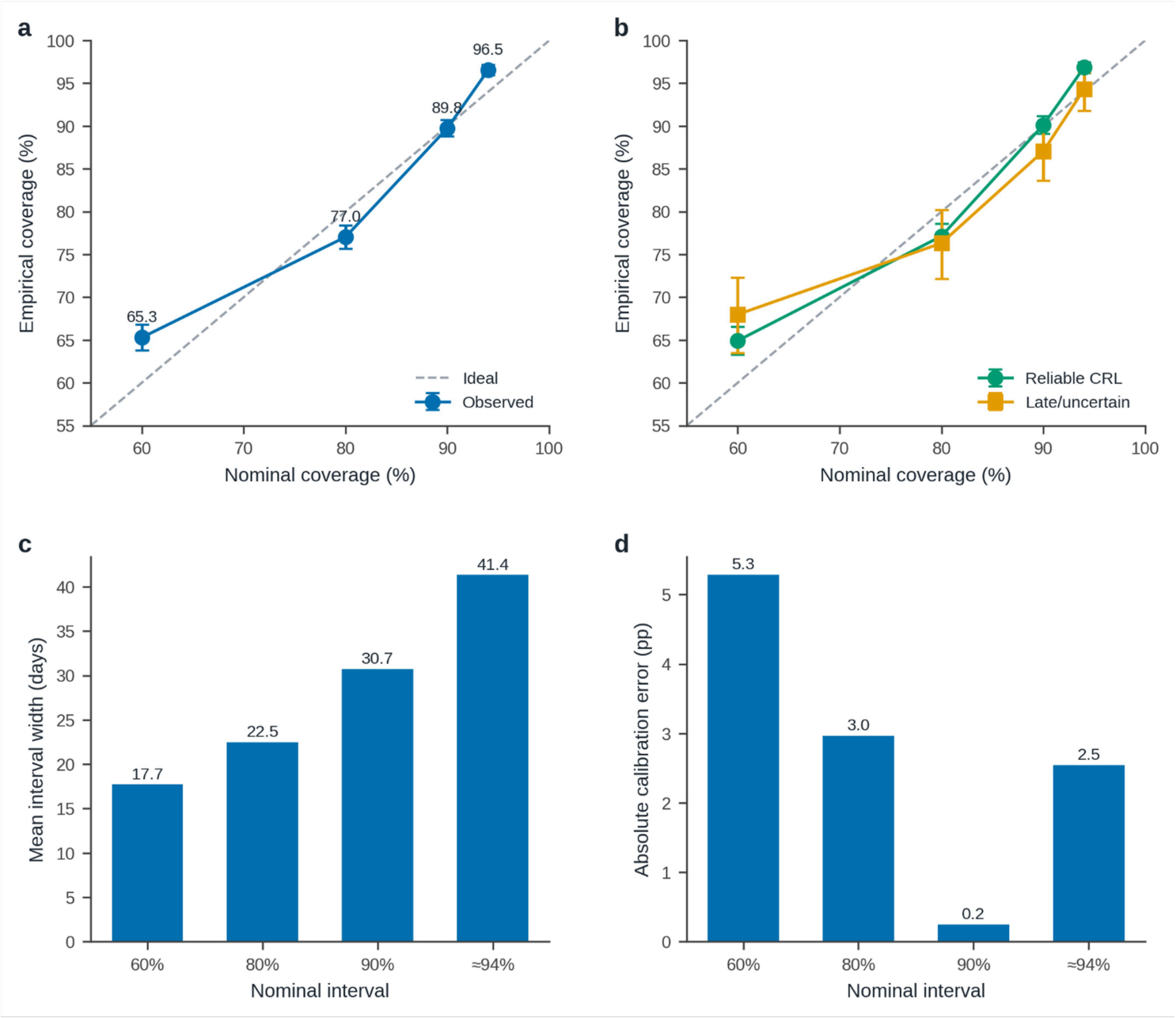


Figure 4

Prediction-interval calibration. (a) Empirical versus nominal coverage of the model's predicted delivery-date quantile intervals in all 3,788 singletons (blue circles and connecting line); the dashed line is the line of perfect calibration, and high-coverage intervals are closest to it. (b) Coverage by dating certainty (green circles, reliable crown–rump length dating; amber squares, late/uncertain dating), showing preserved calibration in both strata. (c) Mean interval width by nominal level. (d) Absolute calibration error (the absolute difference between empirical and nominal coverage) by nominal level. Error bars in a and b are 95% Clopper–Pearson (exact binomial) confidence intervals; panels c and d show point summaries without error bars.

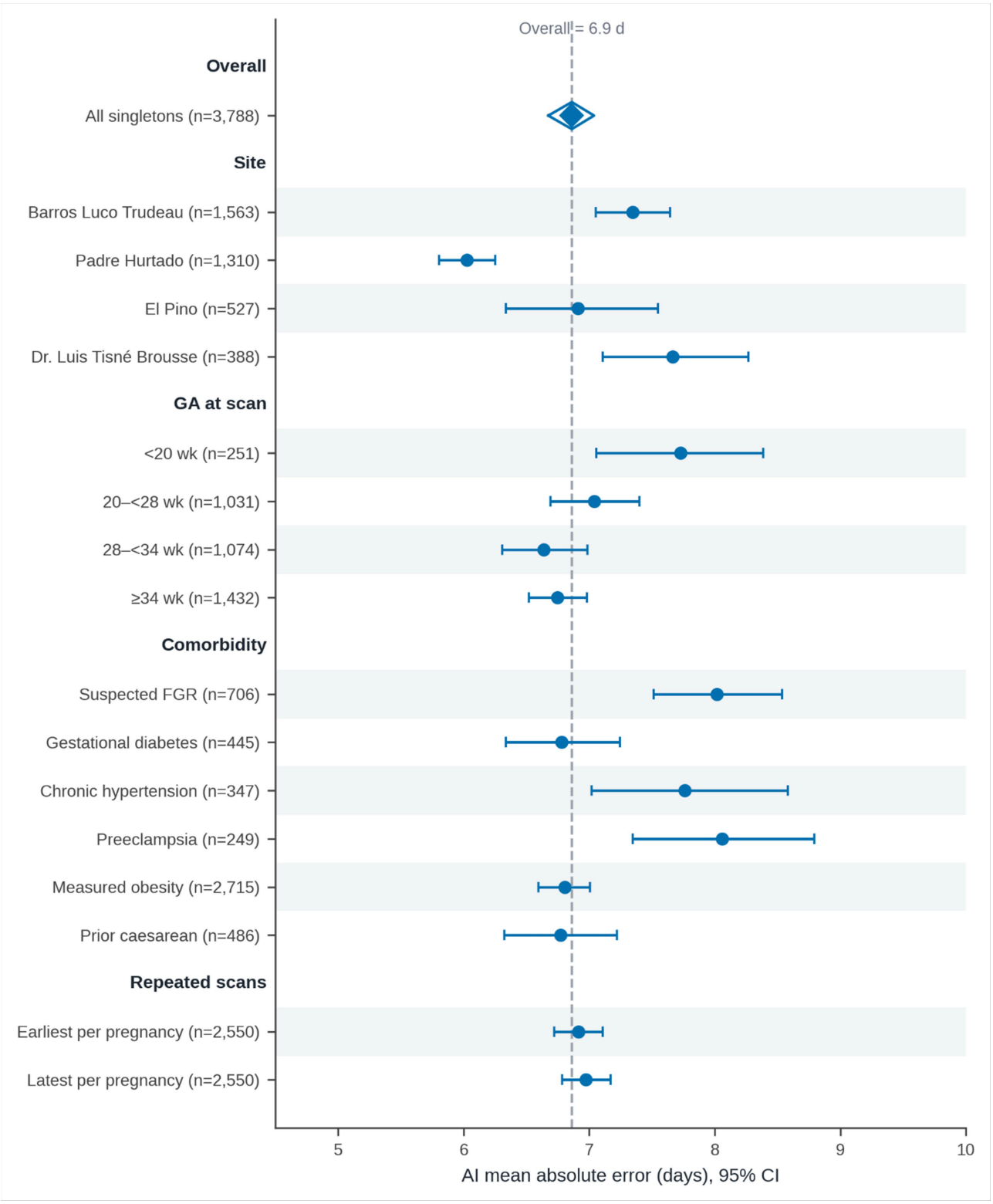


Figure 5

Transportability and robustness. AI mean absolute error (blue circles) by participating hospital, gestational-age-at-scan window, major comorbidity subgroup, and one-scan-per-pregnancy population; the dashed vertical line marks the overall singleton mean absolute error (6.9 days). Group headings are shown in bold and each row is annotated with its sample size. Error bars are 95% confidence intervals

from a non-parametric bootstrap that resampled pregnancies (not individual studies) with replacement (3,000 resamples). Accuracy was consistent across all strata.

Supplementary Files

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

- TRIPODAIChecklist.docx
- SupplementaryInformation.docx