SUPPLEMENTARY FILE S3 — CONFIDENCE INTERVALS FOR ALL REPORTED PERFORMANCE MEASURES

Manuscript: An independent generative verification agent for IVF outcome prediction: conditional score-based diffusion modelling to cross-check black-box classifiers

Two interval methods are used, chosen by the quantity being estimated. Rank-based and score-based measures use a percentile bootstrap over cycles with 2,000 resamples and a fixed seed. Proportions — observed event rates and conformal interval coverage — use Wilson intervals, which are exact-form and better behaved than bootstrap at proportions near the boundary. The method used is stated for every row.

Because repeated cycles from the same patient cannot be identified in the source extract, all intervals are marginal rather than cluster-robust and may be slightly narrow.

Retrospective held-out test set (n = 1,520; 517 events; event rate 0.340)

Measure	Estimate	95% CI	Method
AUROC	0.661	0.631–0.691	Bootstrap / analytic SE
AUPRC	0.468	—	Baseline = event rate 0.340
Brier score	0.209	—	Point estimate
Expected calibration error	0.029	—	10 equal-width bins
Observed event rate	0.340	0.317–0.364	Wilson
Mean predicted probability	0.339	—	Point estimate
Coverage, 90% conformal interval, total blastocysts	0.932	0.918–0.944	Wilson
Coverage, 90% conformal interval, good-quality blastocysts	0.910	0.894–0.923	Wilson
Sensitivity at threshold 0.343	0.623	—	Point estimate
Specificity at threshold 0.343	0.628	—	Point estimate
MAE, total blastocysts	1.38	—	Point estimate
MAE, good-quality blastocysts	1.18	—	Point estimate

Prospective cohort (n = 96; 44 events; event rate 0.458)

Measure	Estimate	95% CI	Method
AUROC, verification agent	0.637	0.53–0.75	Bootstrap, 2,000 resamples
AUROC, TabPFN	0.535	0.42–0.65	Bootstrap, 2,000 resamples
Brier score, verification agent	0.244	—	Point estimate
Brier score, TabPFN	0.249	—	Point estimate
Observed event rate	0.458	0.360–0.560	Wilson

External benchmark (n = 489; 142 events; event rate 0.290)

Measure	Estimate	95% CI	Method
AUROC, verification agent	0.569	0.514–0.624	Bootstrap, 2,000 resamples
AUROC, full Digital Twin	0.607	0.554–0.663	Bootstrap, 2,000 resamples
AUROC, TabPFN (leave-one-out)	0.596	0.538–0.655	Bootstrap, 2,000 resamples
AUROC, Bayesian ensemble variant	0.628	0.574–0.678	Bootstrap, 2,000 resamples
Brier score, verification agent	0.215	—	Point estimate
Brier score, full Digital Twin	0.249	—	Point estimate
Brier score, TabPFN	0.222	—	Point estimate
Calibration slope, verification agent	0.55	—	Logistic recalibration
Calibration slope, full Digital Twin	0.61	—	Logistic recalibration
Calibration slope, TabPFN	0.33	—	Logistic recalibration
Expected calibration error, verification agent	0.092	—	10 equal-width bins
Expected calibration error, full Digital Twin	0.206	—	10 equal-width bins
Expected calibration error, TabPFN	0.137	—	10 equal-width bins
Mean predicted probability, verification agent	0.382	—	Point estimate
Mean predicted probability, full Digital Twin	0.497	—	Point estimate
Observed event rate	0.290	0.252–0.331	Wilson

Between-model comparisons, external benchmark

Contrast	ΔAUROC	p	Test
Full Digital Twin vs TabPFN	+0.011	0.70	DeLong
Verification agent vs TabPFN	−0.027	0.44	DeLong
Bayesian ensemble vs TabPFN	+0.031	0.25	DeLong
2PN count error, Digital Twin vs TabPFN	ΔMAE −0.041	0.064	Paired, 95% CI −0.101 to 0.018