SUPPLEMENTARY FILE S4 — STABILITY ACROSS TRAINING SEEDS

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

Rationale

The Discussion cites evidence that image-based embryo-selection networks give materially different rankings when replicate models are trained differing only in random initialization (Thirumalaraju et al. Fertil Steril 2026;125:277-286). It would be inconsistent to apply that standard to other systems and not to our own. We therefore trained replicate models of the verification agent and evaluated each on the same held-out test set.

Design

Five replicate models were trained. Only the seed governing network initialization, dropout and batch order was varied; every data partition was held fixed at the seed used for the primary analysis, so the replicates differ in training randomness alone and not in which cycles they saw.

Each replicate was evaluated on the identical held-out test set (n = 1,520; 517 events; prevalence 0.3401), which no replicate saw during training.

Setting	Value
Replicates	5 (seeds 1000–1004)
Varied	network initialization, dropout, batch order
Held fixed	all data partitions, test split (seed 42), all hyperparameters
Epochs	200
Synthetic samples per cycle at evaluation	100
Test set	n = 1,520 of 15,193 cycles; prevalence 0.3401

Results

Measure	Median	Min	Max	SD
AUROC	0.6540	0.6472	0.6562	0.0036
AUPRC	0.4617	0.4572	0.4634	0.0025
Brier score	0.2110	0.2104	0.2120	0.0006
Expected calibration error	0.0234	0.0221	0.0342	0.0053

Measure	Median	Min	Max	SD
Sensitivity at optimal threshold	0.6557	0.6286	0.6576	0.0124
Specificity at optimal threshold	0.6012	0.5962	0.6301	0.0137
Optimal threshold	0.3376	0.3371	0.3578	0.0090
Mean predicted probability	0.3522	0.3509	0.3553	0.0018
Observed prevalence	0.3401	0.3401	0.3401	0.0000
Coverage, 90% PI, total blastocysts	0.9296	0.9250	0.9336	0.0033
Coverage, 90% PI, good-quality blastocysts	0.9105	0.9079	0.9539	0.0194
MAE, total blastocysts	1.3033	1.2122	1.3280	0.0453
MAE, good-quality blastocysts	0.9388	0.8431	0.9658	0.0526

Per-replicate values

Seed	AUROC	ECE	Brier	Coverage, Bl	Coverage, good Bl	Minutes
1000	0.6472	0.0342	0.2120	0.934	0.911	53
1001	0.6539	0.0230	0.2111	0.927	0.914	56
1002	0.6540	0.0234	0.2110	0.931	0.908	61
1003	0.6554	0.0301	0.2109	0.925	0.954	62
1004	0.6562	0.0221	0.2104	0.930	0.911	62

Conformal radii, 90% predictive interval

Seed	Total blastocysts	Good-quality blastocysts
1000	3.00	2.00
1001	3.00	2.00
1002	3.00	2.00
1003	3.00	3.00
1004	3.00	2.55
Median	**3.00**	**2.00**

Interpretation

Discrimination varied by 0.009 AUROC across the five replicates. That spread is an order of magnitude smaller than the ranking instability reported for replicate image-based selection networks, and it is also much smaller than the gap between our retrospective estimate and the external one (0.661 versus 0.569). For this architecture, seed variability is not a material source of uncertainty; transportability is.

Interval coverage was stable and close to nominal in every replicate: 92.5% to 93.4% for total blastocysts and 90.8% to 95.4% for good-quality blastocysts, against a nominal 90%. The median of 93.0% and 91.1% reproduces the values obtained for the model reported in the manuscript (93.2% and 91.0%) in five independently trained models.

Relationship to the values reported in the manuscript

The primary analysis reports AUROC 0.661, Brier 0.209 and ECE 0.029 for the model described in the Methods. The replicate median AUROC = 0.654, lies just below the replicate maximum of 0.656 and slightly below the reported value. The replicates were evaluated with 100 synthetic samples per cycle against 200 in the primary analysis; halving the sample count makes each generated median noisier, and the two count medians are inputs to the classifier, so a small loss of discrimination in that direction is expected. We report the comparison rather than reconciling it, since nothing in the manuscript's argument depends on the third decimal of AUROC.