SUPPLEMENTARY FILE S1 — TRIPOD+AI REPORTING CHECKLIST

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

Checklist: TRIPOD+AI (Collins GS, Moons KGM, Dhiman P, et al. BMJ 2024;385:e078378)

Study type: development of a prediction model with internal, prospective and external evaluation — items marked D, E and D;E all apply.

Items that the study does not address are marked **Not reported** with the reason, in accordance with the statement's intent that the checklist document what was and was not done.

Item	Topic	Where addressed
1	Title	Title page. Identifies the study as developing and evaluating a prediction model, the target population (IVF/ICSI cycles) and the predicted outcome.
2	Abstract	Abstract, structured under the journal's seven headings and reporting design, participants, outcome, performance measures with confidence intervals, and a moderated conclusion.
3a	Background: healthcare context and rationale	Introduction, paragraphs 1–2. Prognostic context; rationale for an independent verification agent; references to existing prediction and embryo-selection models.
3b	Target population, intended purpose, intended users	Introduction, paragraph 2 and Discussion, paragraph 1. Intended as a decision-support component used by clinicians and embryologists alongside, not instead of, a predictive model; explicitly not an autonomous decision maker (Methods, "Verification against a black-box predictor").
3c	Known health inequalities between sociodemographic groups	**Not reported.** The source extract contains no ethnicity, socioeconomic or other sociodemographic variables beyond maternal age, so inequalities between such groups could not be described. Noted as a limitation.
4	Objectives	Introduction, final paragraph; Abstract, Objective.

Item	Topic	Where addressed
5a	Sources of data, rationale, representativeness	Methods, "Study design and setting" and "Cohort allocation". Routine clinical records from three reproductive centres for development and internal evaluation; prospective clinical use at one centre; two published public datasets for external evaluation.
5b	Dates of participant data	Methods, "Study design and setting" — 2012 to 2025.
6a	Study setting, number and location of centres	Methods, "Study design and setting"; author affiliations. Three reproductive centres.
6b	Eligibility criteria	Methods, "Participants, inclusion and exclusion criteria".
6c	Treatments received and how handled	Methods, "Study design and setting" and "Outcome definition". Cycles are IVF or ICSI; the cohort mixes fresh and frozen transfers, which is stated and discussed. Stimulation protocol is not a model input.
7	Data pre-processing and quality checking	Methods, "Missing data" and "Model inputs and architecture" (quantile normalisation; post-generation rounding and the constraint that good-quality blastocysts cannot exceed total blastocysts). Pre-processing was applied identically to all cycles.
8a	Outcome definition and time horizon	Methods, "Outcome definition" — clinical pregnancy, fetal cardiac activity on transvaginal ultrasonography 25 days after transfer; defined per transfer.
8b	Qualifications of outcome assessors	**Not reported.** The outcome is an ultrasonographic finding recorded in routine care by the treating clinician; individual assessor characteristics were not retained in the extract.
8c	Blinding of outcome assessment	**Not applicable to the retrospective cohort**, where the outcome was recorded in routine care before any model existed. In the prospective cohort, predictions were generated before embryo culture began and were compared with the outcome only afterwards (Methods, "Cohort allocation"), so outcome ascertainment could not be influenced by the prediction.

Item	Topic	Where addressed
9a	Choice of initial predictors and pre-selection	Methods, "Model inputs and architecture". Seven variables were fixed a priori as those available at the day-1 fertilization check; no data-driven pre-selection was performed.
9b	Definition and measurement of predictors	Methods, "Model inputs and architecture". All are counts or ratios recorded in the routine laboratory record, plus KPIScore, a composite of nine indicators aligned with the Vienna consensus.
9c	Qualifications of predictor assessors	**Not reported.** Predictors are laboratory counts recorded by the embryology team in routine practice; individual assessor characteristics were not retained. Between-embryologist variation is therefore an unmeasured source of variability.
10	How study size was arrived at	Methods, "Cohort allocation". No formal sample-size calculation was performed; the development cohort comprises every consecutive eligible cycle in the source database over the study period, and the evaluation cohorts are of the size available. Stated as a limitation for the prospective cohort (n = 96, 44 events).
11	Missing data handling	Methods, "Missing data". 11 of 15,193 cycles (0.07%) had missing counts, imputed as zero; no cycle was omitted.
12a	How data were used; partitioning	Methods, "Cohort allocation"; Figure 1. Four disjoint partitions plus a test set held out before development.
12b	Handling of predictors: transformation, rescaling	Methods, "Model inputs and architecture" — quantile normalisation to a standard normal distribution with 1,000 quantiles, inverted after generation.
12c	Model type, rationale, building steps, hyperparameter tuning, internal validation	Methods, "Model inputs and architecture" and "Hyperparameters and hyperparameter selection". Conditional score-based diffusion with a Transformer denoiser, a gradient-boosted head and Platt scaling; hyperparameters fixed a priori without automated search, which is stated explicitly; internal validation by a held-out 15% of the diffusion partition.

Item	Topic	Where addressed
12d	Heterogeneity across clusters	**Partially reported.** Between-centre heterogeneity was not modelled: the development extract does not identify the contributing centre per cycle, and no patient identifier is available, so neither centre nor patient clustering could be quantified. The consequence for the confidence intervals is stated in Methods, "Statistical analysis" and in Limitations.
12e	Performance measures and plots, with rationale	Methods, "Statistical analysis"; Figure 2. Discrimination (AUROC, AUPRC), calibration (ECE, calibration slope, observed versus predicted event rate), probabilistic accuracy (Brier score), generative accuracy (MAE, Wasserstein distance, Kolmogorov-Smirnov statistic) and conformal interval coverage.
12f	Model updating arising from evaluation	Discussion, paragraph on the embryo–pregnancy gap. A single leave-one-out linear recalibration of the generated counts was applied post hoc and its effect reported; no recalibration was applied to the probability outputs.
12g	How model predictions were calculated at evaluation	Methods, "Model inputs and architecture" (sample counts per purpose) and "Verification against a black-box predictor" (leave-one-out procedure for the comparator on external data); code in the archived release.
13	Class imbalance	Methods, "Hyperparameters and hyperparameter selection". The gradient-boosted head used class-imbalance weighting; probabilities were subsequently recalibrated by Platt scaling on a held-out split, and calibration after that step is reported (ECE, observed versus predicted event rate).
14	Model fairness	**Not reported.** No fairness analysis was performed. The extract lacks the sociodemographic variables such an analysis would require (see item 3c), and maternal age, the one demographic variable available, is not a model input. This is a limitation for any future deployment and is noted as such.

Item	Topic	Where addressed
15	Model output; classification thresholds	Methods, "Statistical analysis" and Results. Output is a calibrated probability plus conformal predictive intervals for blastocyst counts; the classification threshold maximises the harmonic mean of sensitivity and specificity and is reported with the resulting operating characteristics.
16	Differences between development and evaluation data	Methods, "Cohort allocation" (external harmonisation and reconstructed endpoints) and the paragraph disclosing that ovarian-reserve and body-mass variables are unavailable externally and were held constant; Discussion, paragraph on mechanisms of external miscalibration.
17	Ethical approval and consent	Declarations, Ethics approval.
18a	Funding	Title page and Declarations — no external funding.
18b	Conflicts of interest	Title page and Declarations — none declared.
18c	Protocol	**Not reported.** No study protocol was prepared in advance; this is stated rather than implied.
18d	Registration	**Not registered.** The study is a retrospective model-development study with observational evaluation and was not prospectively registered.
18e	Data sharing	Title page, Data sharing statement. Public benchmark datasets are openly available with DOIs; clinic-level records cannot be shared under confidentiality restrictions, with the reason given.
18f	Code sharing	Title page, Data sharing statement — public repository plus an archived, citable release of the version used here.
19	Patient and public involvement	**No involvement.** Patients and the public were not involved in the design, conduct, reporting, interpretation or dissemination of this study.
20a	Flow of participants	**Figure 1**, with numbers of cycles and outcome events at each stage.
20b	Characteristics overall and by source	**Partially reported.** Cohort sizes, event rates and study dates are given in Methods, Figure 1 and Table 2. A

Item	Topic	Where addressed
		full table of baseline characteristics by data source is not currently included (see note below).
20c	Comparison of predictor distributions, evaluation versus development	**Partially reported.** The Methods state which predictors were unavailable in the external data and were held constant, and the Discussion compares event rates between development and external cohorts. A formal distributional comparison of the shared predictors is not presented.
21	Number of participants and events in each analysis	Methods, "Cohort allocation"; Figure 1; Table 2 (event rate by cohort).
22	Full model specification	Archived code release and public repository; architecture and all hyperparameters in Methods.
23a	Performance estimates with confidence intervals	Results; Table 2; Figure 2. Confidence intervals are reported for discrimination, calibration and interval coverage. Subgroup performance is not reported (see items 3c and 14).
23b	Heterogeneity across clusters	**Not examined**, for the reasons given under item 12d.
24	Results of model updating	Discussion — the effect of the leave-one-out recalibration on count error is reported.
25	Interpretation	Discussion, throughout.
26	Limitations	Discussion, final paragraph, and the dedicated statements on repeated cycles, unavailable external predictors, temporal span, fresh/frozen mixing, absence of a donor-cycle indicator, and the absence of nested cross-validation.
27a	Handling of poor-quality or unavailable input data at implementation	**Partially reported.** The consequences of unavailable inputs are demonstrated empirically in the external evaluation and discussed, but a formal implementation rule for missing or implausible inputs is not specified.
27b	User interaction and expertise required	Methods, "Verification against a black-box predictor" and Discussion, paragraph 1 — the agent returns a flag and an inspectable blastocyst distribution for clinician interpretation, and is not autonomous.

Item	Topic	Where addressed
		Formal user requirements are not specified.
27c	Next steps for future research	Discussion, final paragraph — silent local validation, prespecified recalibration or model updating on one local sample, and prospective retesting on another.