

Online Resource 1

Journal of Assisted Reproduction and Genetics

Latent developmental-arrest burden after cleavage-stage transfer: a counterfactual reanalysis of a published IVF cohort

Supplementary Methods S1–S7, Supplementary Tables S1–S6 and Supplementary Figs. S1–S3. Numbers in square brackets refer to the reference list of the main article; reference [7] is the source cohort report (Dirican EK, Olgan S, Sakinci M, Caglar M. Blastocyst versus cleavage transfers: who benefits? Arch Gynecol Obstet. 2022;305:749–56. <https://doi.org/10.1007/s00404-021-06224-2>).

Supplementary Methods

S1. Continuous modeling of zygote number. Zygote number was modeled as a continuous variable using a natural cubic spline (`patsy cr()` basis) with three and, as a check, four degrees of freedom. The spline basis was interacted with the transfer-day indicator in a logistic model that also contained female age, previous in vitro fertilization (IVF) failures, basal follicle-stimulating hormone (FSH), antral follicle count, cumulus–oocyte complexes, and metaphase II oocytes. The interaction was tested by likelihood ratio against the same model without interaction terms, separately in intention-to-treat (ITT) data (all 1116 cycles) and per-protocol data (1046 cycles, excluding the 70 cancellations). A model with zygote number entered linearly and a model with the $\leq 5/\geq 6$ dichotomy are reported for comparison (Supplementary Table S2). Predicted clinical pregnancy rates were obtained at the mean of every covariate (Supplementary Fig. S2A) and, as a check on that convention, by marginal standardization over the observed covariate distribution: for each value of the zygote count in turn, that value was assigned to every cycle in the dataset, predictions were formed under an “everyone day 3” and an “everyone day 5” assignment, and the predictions were averaged over the sample (Supplementary Table S5; Supplementary Fig. S2B). Confidence intervals (CIs) for the marginally standardized differences come from 2000 nonparametric bootstrap replications of the whole dataset with the model refitted in each replicate (seed 42). Predictions are displayed for 1–10 zygotes, which covers 1081 of the 1116 cycles; the spline extrapolates unstably beyond 10 zygotes, where only 35 cycles are observed, and those values are given in Supplementary Table S5 for completeness only.

S2. Multiplicity in the threshold scan. Six candidate dichotomies (≥ 3 to ≥ 8 zygotes) were each tested for interaction with transfer day. Because these six tests address the same underlying question, *P* values are additionally reported after Holm adjustment (Supplementary Table S1). No correction was applied to the primary analyses, which were not selected from a scan: the continuous spline model requires no cutoff, and the $\leq 5/\geq 6$ dichotomy was inherited from the previously published analysis of this cohort [7] rather than chosen here.

S3. Covariate-adjusted latent-arrest estimate (g-computation). Day-5 cancellation was modeled by logistic regression on zygote number, female age, previous IVF failures, basal FSH, antral follicle count, cumulus–oocyte complexes, and metaphase II oocytes, fitted in the 532 day-5 cycles. All continuous terms were entered linearly. The fitted model was then used to predict a counterfactual cancellation probability for each of the 584 day-3 cycles, and these probabilities were summed to give an adjusted expected latent-arrest count. Confidence intervals came from 5000 bootstrap replications in which the two arms were resampled independently and the model was refitted within each replicate.

Model diagnostics. The dataset was complete, so no observations were dropped and no imputation was used. The model contained 8 parameters and was fitted to 70 cancellation events (8.8 events per parameter); it converged in 6 iterations, and the largest absolute coefficient (1.47) and largest standard error (1.17) gave no indication of separation. None of the 5000 bootstrap replicates failed to converge. Entering zygote number and age as natural cubic splines instead of linear terms did not improve fit (likelihood-ratio $\chi^2 = 4.14$; *df* = 2; *P* = 0.126), supporting the linearity assumption. Apparent discrimination was moderate (*c*-statistic, 0.782; Brier score, 0.097). On 10-fold cross-validation the *c*-statistic fell to 0.749, the Brier score rose to 0.101, and the calibration slope was 0.818 (95% CI, 0.560–1.077) with an intercept of -0.285 (95% CI, -0.780 to 0.210), indicating mild optimism, although the slope interval includes one. A decile-based Hosmer–Lemeshow test on the apparent predictions indicated

some departure from perfect calibration ($\chi^2 = 20.43$; $df = 8$; $P = 0.009$). The covariate-adjusted estimate is therefore presented as a sensitivity analysis supporting the zygote-only standardization, not as a definitive adjusted estimate.

S4. Bootstrap confidence intervals across ϕ . Within each of 5000 stratified bootstrap replicates the residual difference was recomputed at every value of ϕ from 0% to 30% in steps of 5 percentage points, so the intervals in Supplementary Table S3 and the bands in Supplementary Fig. S3 derive from the same resamples as the primary $\phi = 0$ interval. At $\phi = 0$ the procedure reproduces the intervals reported in Table 2 exactly.

S5. Confidence interval for the latent-arrest burden. The latent-arrest count is $A = \sum n_3(z) \cdot \hat{p}_5(z)$, where $n_3(z)$ is the number of day-3 cycles at zygote count z and $\hat{p}_5(z) = c_5(z)/n_5(z)$ is the day-5 cancellation fraction observed at that count. Its uncertainty arises from the binomial sampling of the $c_5(z)$. Each $c_5(z)$ was redrawn from $\text{Binomial}(n_5(z), \hat{p}_5(z))$ and A recomputed, 20,000 times (seed 42), and the 2.5th and 97.5th percentiles taken. This gives 22.1% (95% CI, 17.3 to 26.8) of day-3 cycles with ≤ 5 zygotes, 4.1% (1.2 to 7.6) with ≥ 6 , and 18.6% (14.7 to 22.5) overall. A delta-method interval from $\text{Var}(A) = \sum n_3(z)^2 \cdot \hat{p}_5(z) \{1 - \hat{p}_5(z)\} / n_5(z)$ agrees to within 0.5 percentage points in every stratum, and both agree with the covariate-adjusted g-computation intervals of S3 (21.8%, 16.9 to 27.0; 2.5%, 0.8 to 6.2; 18.1%, 14.2 to 22.3), which are estimated by a different route. These intervals reflect sampling variability in the observed cancellation rates only. They do not cover the exchangeability assumption by which those rates are transported to the day-3 arm, which is not testable in these data and is discussed in the main text. Zygote counts at which no day-5 cancellation was observed (nine or more) contribute exactly zero to A under this estimator; no shrinkage or continuity correction was applied. The calculation above is reproduced by `scripts/18_burden_interval.py` in Online Resource 2, which uses only the published counts.

S6. Consistency of the archived analytic dataset with the original cohort report. Because this is a secondary analysis of a cohort that has already been published, the archived analytic file was checked against every count reported in the original article before any new analysis was run [7]. Of the 45 published quantities that can be recomputed from the archived file, 42 reproduce exactly (`scripts/15_verify_against_published.py` in Online Resource 2). The arm sizes (584 and 532), the number of extended-culture cancellations (70), the clinical pregnancy counts at each zygote number from 1 to 9 or more, the stratum-specific cancellation counts (64/383 for ≤ 5 zygotes; 6/149 for ≥ 6 zygotes), and the baseline means for age, basal FSH, antral follicle count, cumulus-oocyte complexes, and metaphase II oocytes all reproduce the published values exactly (Supplementary Table S6). Refitting the original adjusted logistic model on the archived file reproduces its ≤ 5 -zygote estimate to three decimal places (adjusted odds ratio [aOR], 0.874; 95% CI, 0.635–1.204; identical to the published value) and gives 3.27 (95% CI, 1.83–5.83) for the ≥ 6 -zygote stratum, where the original reported 3.12 (95% CI, 1.80–5.43). That difference is small, leaves the direction and significance of the contrast unchanged, and is confined to the adjusted models; the two reports used different software. Its most likely source is the previous-IVF-failure covariate, whose overall means in the archived file (0.51 in the day-3 arm and 0.40 in the day-5 arm) differ slightly from the published Table 1 (0.52 and 0.42) while matching the ≤ 5 -zygote stratum of the archived file exactly (0.52 and 0.42). No quantity entering the counterfactual decomposition depends on that covariate: the decomposition uses only the allocation, the zygote count, the cancellation indicator, and the pregnancy indicator, all of which reproduce the published tables exactly.

One further discrepancy is recorded here because the verification log ships with this submission. In Figure 1 of the source report the two-pronuclear-zygote totals appear on the branches as day 3 = 2337 and day 5 = 2,088, the reverse of the archived analytic dataset (day 3 = 2088; day 5 = 2337). The archived orientation is the internally consistent one: $2088/584 = 3.58$ and $2337/532 = 4.39$ are the mean zygote counts printed in Table 1 of the present report and in the source article. No quantity used here depends on those totals and they are not among the 45 values checked above, but the apparent transposition is noted so that a reader who opens `outputs/15_verify_against_published.log` is not left to discover it unexplained.

S7. Which quantities are re-reported and which are new. The source report [7] reported the cohort construction and flow, the baseline characteristics of the two arms, clinical pregnancy at each zygote count and in each stratum for both arms, the overall day-5 cancellation rate of 13.2% with its stratum split of 16.7% versus 4.0%, adjusted odds ratios for transfer day of 0.874 with ≤ 5 zygotes and 3.122 with ≥ 6 , and

implantation, cryopreservation and embryo-utilization rates. New in the present report are the cancellation rate at each individual zygote count rather than by stratum; the counterfactual application of those rates to the day-3 arm and the latent-arrest and putatively futile-transfer estimates that follow, with their intervals; the implied day-3-progressable subgroup and its residual difference against the day-5 per-protocol rate; the per-protocol analysis, which the source article deliberately did not perform; the continuous spline model with marginal standardization; the covariate-adjusted g-computation; the ϕ sensitivity analysis; the post hoc equivalence testing; the standardized mean differences; the threshold scan with Holm adjustment; and the verification of the archived dataset against every count printed in the source article. Table 1, Table 2 and the legend of Figure 2 of the main text mark which values were previously published.

Supplementary Tables

Supplementary Table S1

Sensitivity of the transfer-day-by-zygote-group interaction to the dichotomization threshold (intention-to-treat model), with multiplicity adjustment.

Threshold	Interaction coefficient	95% CI	LR <i>P</i> value	Holm-adjusted <i>P</i>	AIC
≥3 zygotes	+0.49	−0.15 to +1.12	0.129	0.258	1297.1
≥4 zygotes	+0.35	−0.20 to +0.90	0.213	0.258	1311.4
≥5 zygotes	+0.90	+0.32 to +1.47	0.002	0.008	1305.6
≥6 zygotes (used for presentation)	+1.25	+0.61 to +1.89	0.0001	0.0006	1303.0
≥7 zygotes	+1.29	+0.59 to +2.00	0.0002	0.0012	1305.7
≥8 zygotes	+1.00	+0.18 to +1.82	0.015	0.044	1312.5

Note: Intention-to-treat logistic regression adjusted for female age, previous IVF failures, basal FSH, antral follicle count, cumulus–oocyte complexes, and metaphase II oocytes. All six analyses were exploratory. The lowest AIC occurred at the ≥3-zygote cutoff (1297.1), where the interaction was not significant; AIC therefore did not identify a clinically meaningful single best threshold and was not the basis for any selection. The ≤5/≥6 categorization was not selected here: it is inherited from the previously published analysis of the same cohort [7], where it was itself reported as a post hoc subgroup analysis, and was retained unchanged for comparability rather than as a prespecified threshold. It is not offered as a validated treatment-decision cutoff. At the ≥6-zygote threshold the per-protocol interaction was also significant (likelihood-ratio $\chi^2 = 11.16$; *df* = 1; *P* < 0.001).

AIC = Akaike information criterion; CI = confidence interval; FSH = follicle-stimulating hormone; IVF = in vitro fertilization; LR = likelihood ratio.

Supplementary Table S2

Interaction between transfer day and zygote number, with zygote number modeled continuously and dichotomously.

Model of zygote number	Data	LR χ^2 (df)	<i>P</i>	AIC with interaction	AIC without interaction
Natural cubic spline, 3 df	ITT	10.93 (2)	0.004	1297.6	1304.5
Natural cubic spline, 4 df	ITT	12.39 (3)	0.006	1294.0	1300.4
Linear	ITT	7.79 (1)	0.005	1303.9	1309.7
Dichotomized at ≥6	ITT	15.24 (1)	<0.001	1303.0	—
Natural cubic spline, 3 df	Per protocol	7.29 (2)	0.026	1256.2	1259.4
Natural cubic spline, 4 df	Per protocol	9.43 (3)	0.024	1253.5	1256.9
Linear	Per protocol	3.79 (1)	0.051	1260.7	1262.5
Dichotomized at ≥6	Per protocol	11.16 (1)	<0.001	1256.9	—

Note: All models were adjusted for female age, previous IVF failures, basal FSH, antral follicle count, cumulus–oocyte complexes, and metaphase II oocytes. In the linear model the interaction corresponded to an odds ratio of 1.149 per additional zygote (95% CI, 1.041–1.268) under ITT and 1.104 (95% CI, 0.999–1.220) per protocol.

Model-predicted day-5-minus-day-3 clinical pregnancy rate at covariate means (ITT; natural cubic spline, 3 df), by zygote number: 1 zygote, –3.8 pp; 2, –3.0; 3, –1.8; 4, +0.1; 5, +2.7; 6, +6.1; 7, +10.3; 8, +15.4; 9, +21.1; 10, +27.0. The difference therefore increased monotonically across the displayed range of 1–10 zygotes and crossed zero at approximately four zygotes. The observed range of the data is 1–15 zygotes; see Supplementary Table S5.

AIC = Akaike information criterion; CI = confidence interval; FSH = follicle-stimulating hormone; ITT = intention-to-treat; IVF = in vitro fertilization; LR = likelihood ratio; pp = percentage points.

Supplementary Table S3

Residual difference across ϕ , with bootstrap 95% confidence intervals at every value of ϕ .

ϕ (%)	≤ 5 zygotes: difference (95% CI)	≥ 6 zygotes: difference (95% CI)	Overall: difference (95% CI)
0	–1.3 (–8.6 to +5.7)	+23.6 (+11.7 to +35.4)	+5.9 (–0.6 to +12.0)
5	+0.1 (–7.1 to +7.0)	+23.8 (+12.0 to +35.6)	+7.0 (+0.6 to +13.0)
10	+1.5 (–5.6 to +8.5)	+24.0 (+12.2 to +35.8)	+8.2 (+1.9 to +14.2)
15	+2.9 (–4.2 to +9.8)	+24.2 (+12.4 to +36.0)	+9.3 (+3.1 to +15.3)
20	+4.3 (–2.8 to +11.2)	+24.4 (+12.7 to +36.2)	+10.5 (+4.3 to +16.4)
25	+5.7 (–1.4 to +12.6)	+24.7 (+12.9 to +36.5)	+11.6 (+5.5 to +17.6)
30	+7.2 (+0.0 to +14.1)	+24.9 (+13.1 to +36.7)	+12.7 (+6.6 to +18.7)

Note: Differences are in percentage points: day-5 per-protocol clinical pregnancy rate minus the implied day-3 progressable rate, the latter recomputed as $(P - \phi A)/(N - A)$, where P is the observed number of day-3 pregnancies, A is the expected number of latent-arrest cycles, and N is the number of day-3 cycles. ϕ is the assumed probability of clinical pregnancy in a day-3 cycle belonging to the latent-arrest stratum. Confidence intervals are from 5000 stratified bootstrap replications, with the difference recomputed at every ϕ within each replicate. In the ≥ 6 -zygote stratum the interval excluded zero at every ϕ examined. In the ≤ 5 -zygote stratum the lower bound remained above –10 percentage points throughout but fell below –7 points at $\phi \leq 5\%$. At $\phi = 30\%$ the lower bound was +0.02 and therefore excluded zero. No noninferiority margin was prespecified for this analysis.

CI = confidence interval.

Supplementary Table S4

Exploratory residual difference by age and zygote stratum.

Age	Zygote stratum	Day 3, n (pregnancies)	Day 5 per protocol, n (pregnancies)	Difference (pp)	95% CI
<35 y	≤ 5 zygotes	229 (69)	203 (67)	–2.3	–11.8 to +6.9
<35 y	≥ 6 zygotes	83 (25)	94 (49)	+20.7	+4.8 to +34.8
≥ 35 y	≤ 5 zygotes	242 (51)	116 (33)	+2.1	–8.0 to +12.4
≥ 35 y	≥ 6 zygotes	30 (4)	49 (23)	+33.1	+14.8 to +49.8

Note: The three-way age-by-zygote-group-by-transfer-day interaction was not significant (likelihood-ratio $\chi^2 = 0.78$; df = 1; $P = 0.38$), so this analysis is hypothesis-generating only. The ≥ 35 -year, ≥ 6 -zygote cell rests on small numbers. Bootstrap 95% confidence intervals, 5000 replications.

CI = confidence interval; pp = percentage points.

Supplementary Table S5

Marginally standardized clinical pregnancy rate by zygote number, with cycle counts.

2PN zygotes	Cycles (n)	Day 3 / day 5 (n)	Marginal CPR, day 3 (%)	Marginal CPR, day 5 (%)	Difference (pp)	95% CI
1	198	139 / 59	20.7	16.9	-3.8	-11.5 to +4.2
2	218	131 / 87	24.0	21.0	-3.0	-9.7 to +3.2
3	177	88 / 89	27.4	25.6	-1.8	-8.1 to +4.3
4	151	69 / 82	30.4	30.5	+0.1	-7.1 to +6.7
5	110	44 / 66	32.8	35.5	+2.6	-6.3 to +10.8
6	71	24 / 47	34.1	40.1	+6.0	-4.2 to +15.6
7	60	28 / 32	33.9	44.1	+10.2	-0.4 to +20.7
8	46	22 / 24	31.9	47.1	+15.2	+4.1 to +26.7
9	29	10 / 19	28.2	49.0	+20.8	+8.8 to +34.7
10	21	12 / 9	23.3	50.0	+26.7	+11.5 to +43.7
11–15 ^a	35	17 / 18	—	—	—	—

Note: Estimates come from the same intention-to-treat logistic model as Supplementary Table S2 (natural cubic spline, 3 df, in zygote number interacted with transfer day, adjusted for female age, previous IVF failures, basal FSH, antral follicle count, cumulus–oocyte complexes, and metaphase II oocytes). For each row the zygote count was set to that value for every cycle in the dataset and the predicted probability was averaged over the observed covariate distribution, first under an “everyone day 3” and then under an “everyone day 5” assignment. Bootstrap 95% confidence intervals are from 2000 replications of the whole dataset with the model refitted in each replicate (seed 42). The point estimates agree closely with the predictions at covariate means reported in Supplementary Table S2 (for example, -3.8 vs. -3.8 pp at one zygote and +26.7 vs. +27.0 pp at 10 zygotes).

^a Zygote counts of 11–15 are pooled here because only 35 cycles fall in that range and the spline extrapolates unstably above 10 zygotes; per-count estimates in that range are reported in Online Resource 2 (outputs/out_marginal_standardised_zygote.csv) but are not interpretable and are not plotted.

CI = confidence interval; CPR = clinical pregnancy rate; FSH = follicle-stimulating hormone; IVF = in vitro fertilization; pp = percentage points; 2PN = two-pronuclear.

Supplementary Table S6

Clinical pregnancy and cancellation counts as published in the original cohort report and as recomputed from the archived analytic dataset.

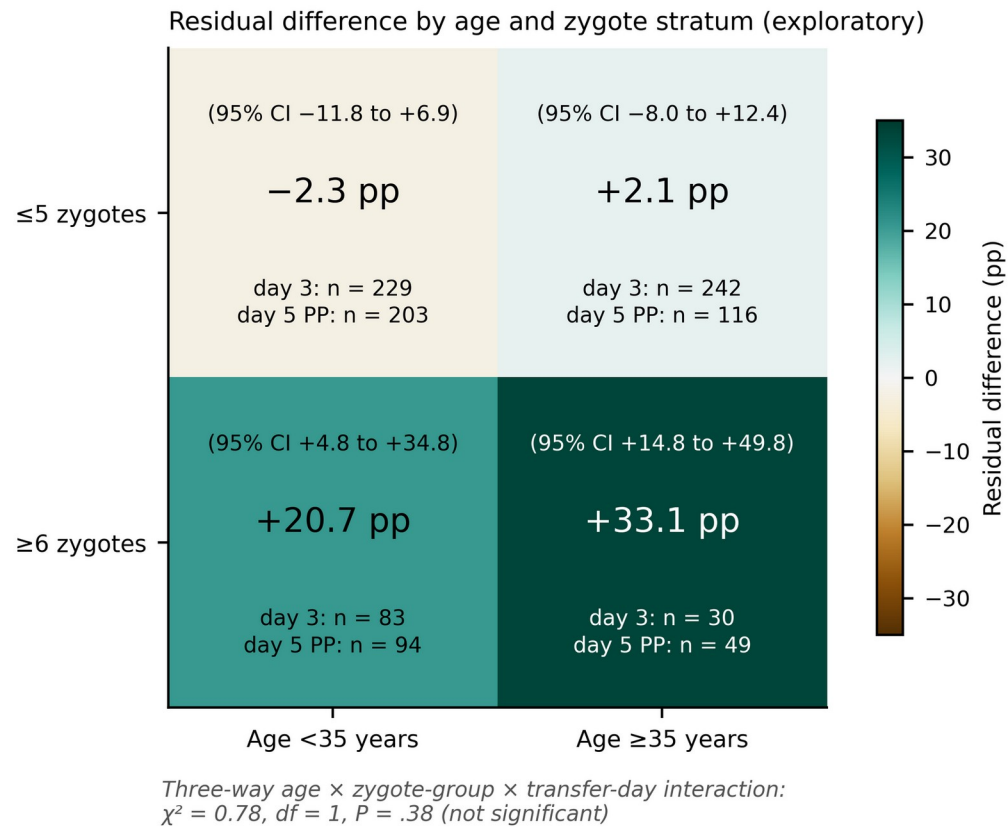
2PN zygotes	Day 3, published	Day 3, recomputed	Day 5, published	Day 5, recomputed
1	19/139	19/139	7/59	7/59
2	30/131	30/131	16/87	16/87

2PN zygotes	Day 3, published	Day 3, recomputed	Day 5, published	Day 5, recomputed
3	28/88	28/88	31/89	31/89
4	27/69	27/69	22/82	22/82
5	16/44	16/44	24/66	24/66
6	7/24	7/24	20/47	20/47
7	7/28	7/28	20/32	20/32
8	4/22	4/22	9/24	9/24
≥9	11/39	11/39	23/46	23/46
≤5 (stratum)	120/471	120/471	100/383	100/383
≥6 (stratum)	29/113	29/113	72/149	72/149
All	149/584	149/584	172/532	172/532
Cancellations, ≤5	—	—	64/383	64/383
Cancellations, ≥6	—	—	6/149	6/149

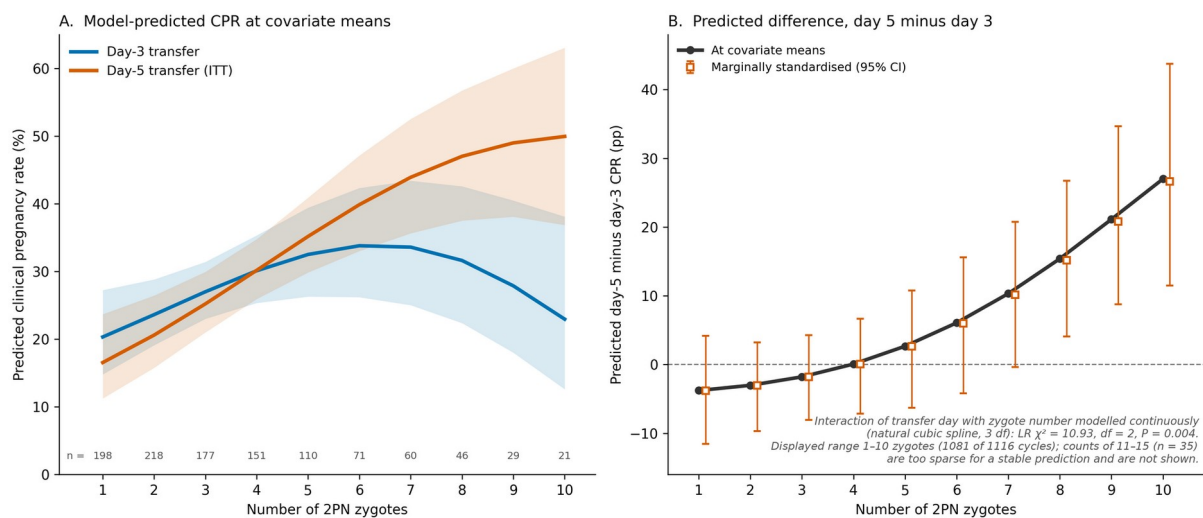
Note: Published values are from the original cohort report [7], Table 2 and Results. Recomputed values are from the archived analytic dataset used for the present analysis (Online Resource 2, outputs/). Entries are pregnancies per cycles. Every entry agrees.

2PN = two-pronuclear.

Supplementary Figures

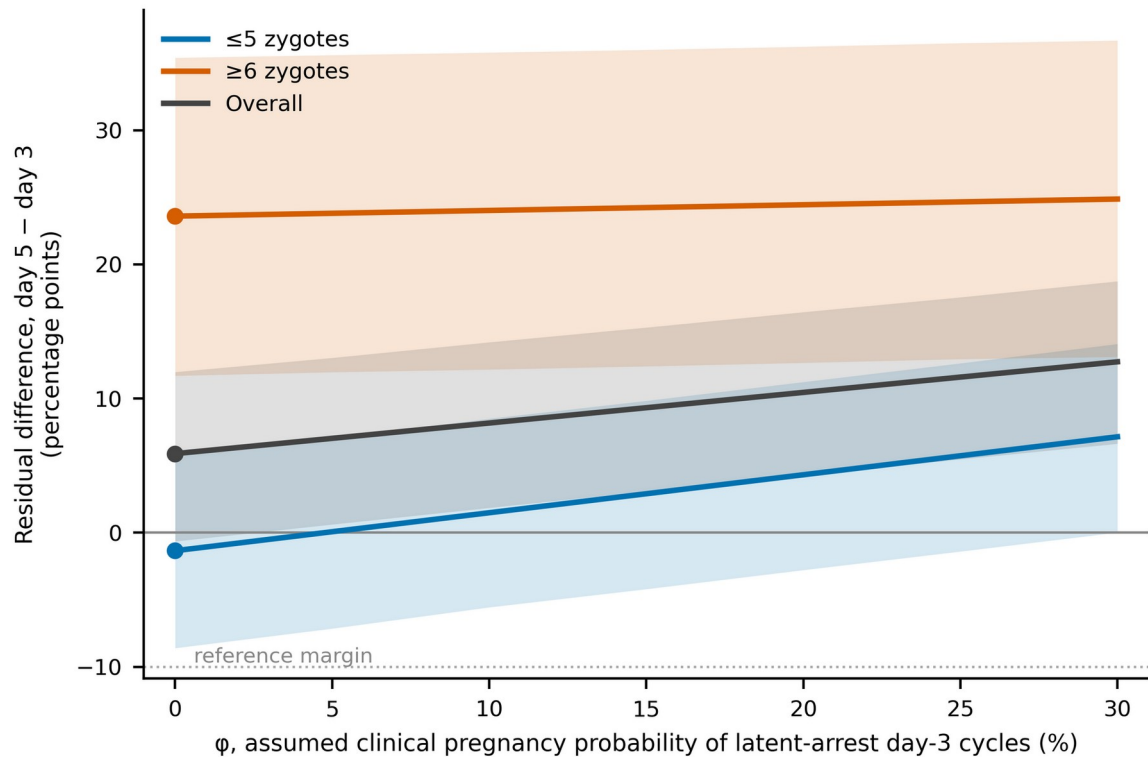


Supplementary Fig. S1 Exploratory heat map of the residual difference (percentage points) by age and zygote stratum, on a colorblind-safe diverging scale (brown, negative difference; teal-green, positive; centered at zero), with the point estimate, bootstrap 95% confidence interval, and the day-3 and day-5 per-protocol cell sizes printed in each cell. The three-way age-by-zygote-group-by-transfer-day interaction was not significant (likelihood-ratio $\chi^2 = 0.78$; $df = 1$; $P = 0.38$), so the panel is descriptive and hypothesis-generating.



Supplementary Fig. S2 Zygote number modeled as a continuous variable. **(A)** Clinical pregnancy rate predicted by a logistic model with a natural cubic spline (3 df) in zygote number interacted with transfer day and adjusted for female age, previous in vitro fertilization failures, basal follicle-stimulating hormone, antral follicle count, cumulus–oocyte complexes, and metaphase II oocytes, evaluated at the mean of every covariate; blue, day-3 transfer; vermilion, day-5 transfer (intention-to-treat); shaded bands are 95%

confidence intervals. **(B)** The predicted day-5-minus-day-3 difference from the same model (black circles and line, evaluated at covariate means) together with the corresponding marginally standardized difference and its bootstrap 95% confidence interval (open vermilion squares), which averages the prediction over the observed covariate distribution rather than evaluating it at the means. The two agree closely. The difference increased monotonically across the displayed range of 1–10 zygotes and crossed zero at approximately four zygotes. The displayed range covers 1081 of the 1116 cycles; the number of cycles contributing at each zygote count is printed along the axis of panel A, and the 35 cycles with 11–15 zygotes are too sparse for a stable prediction and are not displayed (Supplementary Table S5). The interaction was significant (likelihood-ratio $\chi^2 = 10.93$; $df = 2$; $P = 0.004$), indicating a graded relationship rather than a discrete threshold.



Supplementary Fig. S3 Sensitivity of the residual difference to ϕ , the assumed clinical pregnancy probability in day-3 cycles assigned to the latent-arrest stratum. Thick lines are point estimates, and shaded bands are bootstrap 95% confidence intervals from 5000 stratified replications at each value of ϕ . The filled marker denotes the primary specification of $\phi = 0$. The dotted line at -10 percentage points is a reference margin only and was not a prespecified noninferiority margin.