Skepthical review: *Characterizing the Variability and Correlates of U.S. ART Clinic Performance During the COVID-19 Pandemic (2020-2022)*

Summary

The manuscript studies year-to-year variability in U.S. ART clinic performance during 2020–2022 using CDC NASS clinic-level aggregates, organized as clinic × stratum panels (egg source × age group for own eggs). For each clinic–stratum with ≥ 2 years of data, the paper computes variability measures (SD and coefficient of variation, CV) for success metrics (e.g., % live births per intended/actual retrieval; donor-egg live-birth rate) and efficiency metrics (e.g., average transfers per intended retrieval; intended retrievals per live birth) (Secs. 2.2–2.4). The authors then relate these variability measures to a clinic "volume" measure and geography (state) using Spearman correlations, Kruskal–Wallis tests, and OLS regressions with state fixed effects (Secs. 2.5, 3.3–3.4).

Empirically, the manuscript reports substantial observed variability in several outcomes and a central, counterintuitive association: higher-volume clinic strata show greater variability in success-rate metrics (especially when summarized by CV), while some efficiency metrics exhibit lower variability with higher volume (Sec. 3.3–3.4.1). The paper also documents important data constraints, including (i) inability to map "% live birth per transfer" for 2020–2022 and (ii) an apparent 2022 donor-egg reporting anomaly (0% live-birth rates for all clinics) that directly contaminates SD/CV computations for donor-egg outcomes (Secs. 3.1–3.2, 3.5).

Overall, the descriptive objective is valuable—variability (not just average performance) is policy-relevant for surveillance and quality measurement. However, several conceptual and statistical issues currently limit interpretability: the analysis does not separate true performance instability from expected sampling noise given denominators; variability measures are computed from only 2–3 annual points and can be unstable (especially CV with small means/zeros); "volume" is defined inconsistently across Methods/Results; donor-egg conclusions are not reliable given the 2022 anomaly; and OLS inference is weak for highly skewed, nonnegative outcomes with evident assumption violations (Secs. 2.4–2.5.4, 3.4.3, 3.5). Addressing these points (and improving metric extraction documentation and figure consistency) would substantially strengthen the credibility and usefulness of the findings.

Strengths

- Timely and policy-relevant focus on stability/variability of clinic-reported ART outcomes during 2020–2022, a period of system disruption (Sec. 1).
- Uses a large near-census national surveillance dataset (NASS) and a clear clinic × stratum panel construction (Secs. 2.1–2.3, 3.1).

- Examines multiple outcome families (success and resource-use/efficiency proxies) and reports both SD and CV, which helps reveal scale-dependent vs scale-normalized dispersion (Secs. 2.4, 3.3).
- Employs complementary association tools (Spearman correlations, Kruskal–Wallis, OLS with state fixed effects) and includes regression diagnostic plots acknowledging modeling limitations (Secs. 2.5, 3.4.3).
- Transparent about key data limitations (unmappable live-birth-per-transfer metric; donor-egg 2022 anomaly), which is crucial for responsible interpretation (Secs. 3.1–3.2, 3.5).
- Figures generally attempt to make complex relationships (variability distributions; variability–volume patterns; geographic differences) accessible and are a good basis for a stronger final presentation once labeling/consistency issues are resolved (Secs. 3.3–3.4; Figs. 2–17).

Major issues

1. **Core conceptual ambiguity: the paper interprets "performance variability" as reflecting clinic instability/disruption (implicitly resilience), but the measured SD/CV of reported annual rates conflates (i) true underlying changes in clinic quality/processes, (ii) within-stratum case-mix shifts, (iii) expected sampling variability of rates given finite denominators (cycle counts), and (iv) reporting artifacts (clearly present for donor eggs in 2022) (Secs. 1, 3.5, 4). Without explicitly modeling denominator-driven uncertainty, the analysis risks labeling statistical noise (or changing composition) as "volatile performance," and the counterintuitive volume–variability finding is especially hard to interpret under this conflation (Secs. 3.3–3.4.1).**

   Recommendation: Add an explicit estimand/interpretation paragraph (Sec. 1 and/or Sec. 2.4): clarify whether the goal is (a) observed volatility in published clinic metrics (consumer-facing) or (b) latent instability in underlying clinic performance. If aiming at (b), incorporate denominators/precision into the analysis. Concretely, in Sec. 2.4–2.5 and results (Sec. 3.3–3.4), implement at least one of: (i) compute an "excess variability" measure by comparing observed across-year variance of a rate to its expected binomial sampling variance $p(1-p)/N$ using stratum-specific denominators; (ii) shrink annual rates toward a clinic–stratum mean (empirical Bayes/meta-analytic) before computing across-year variability; or (iii) model annual counts (e.g., live births) as binomial with denominators (retrievals/transfers) and estimate a time-varying clinic component. At minimum, add sensitivity analyses restricting to clinic–strata with sufficiently large denominators in each year (e.g., $N \geq 25/50$) and show whether the volume–variability association persists.

2. **Pandemic-era / COVID-19 framing is not operationalized: the manuscript analyzes only 2020–2022 and contains suggestive language about "pandemic disruption," but provides no pre-pandemic baseline and no direct/proxy measures of local pandemic intensity (shutdown timing, case rates, policy) (Sec. 1; Secs. 3.3–3.5, 4). As written, attribution of observed patterns to COVID-era disruption is speculative because similar variability could exist in non-pandemic years.**

   Recommendation: Revise Sec. 1, Sec. 3.5, and Sec. 4 to frame the study as describing variability during 2020–2022 rather than identifying pandemic effects, unless additional analyses are added. If feasible, extend extraction to include at least one pre-pandemic window (e.g., 2017–2019) and compare variability distributions and volume–variability relationships pre vs. during 2020–2022; describe methods in Sec. 2.4–2.5 and report in Sec. 3.3–3.5. Alternatively, link external COVID intensity indicators at the state/county level and test whether variability is higher in higher-intensity areas.

3. **Clinic volume is defined inconsistently across Methods and Results, undermining the central volume–variability findings. Sec. 2.4 defines Avg_Clinic_Volume as the mean Cycle_Count across years within a clinic–stratum (possibly only for years with metric data), while Sec. 3.1 refers to a "maximum" Stratum_Cycle_Count and later reverts to averages (Secs. 2.4 vs. 3.1 vs. 3.4.1). It is also unclear how zeros, missing years, duplicate rows, and multi-stratum clinics are handled; and stratum-specific volume may conflate "size" with case-mix (age/egg-source composition).**

   Recommendation: In Secs. 2.2–2.4, provide a single precise mathematical definition for: (i) per-year stratum volume; (ii) Stratum_Cycle_Count (if used); and (iii) Avg_Clinic_Volume, including the averaging set (all three years vs only years with non-missing outcomes) and treatment of zero-cycle years. Resolve the Sec. 3.1 "maximum" vs "average" inconsistency and ensure all figures/models use the same definition. Add a sensitivity analysis using an overall clinic-level volume (total cycles across strata) alongside stratum-specific volume, or explicitly limit conclusions to stratum-specific volume (Sec. 3.5, Sec. 4). Also clarify whether volume quartiles (Sec. 3.4.1; Figs. 10–17) are computed within stratum or globally.

4. **Donor-egg results are not reliable because the manuscript documents an apparent systemic anomaly in 2022 donor-egg live-birth rates (0% for all clinics), yet donor-egg SD/CV, correlations, and regressions appear to include 2022 (Secs. 3.2–3.4.1, 3.5). With one year mechanically set to zero, across-year variability becomes largely an artifact of 2020–2021 values and the mean, distorting any donor-egg volume/geography associations.**

   Recommendation: Pre-specify and implement donor-egg analytic scenarios in Sec. 2.4–2.5: (i) treat 2022 donor-egg outcome values as missing; and/or (ii) restrict donor-egg variability analyses to 2020–2021 (noting $n = 2$ limitations); and/or (iii) omit donor-

egg variability analyses until the anomaly is resolved. Recompute and report donor-egg descriptive/association results under the anomaly-robust scenario(s) (Sec. 3.3–3.4) and revise wording in the Abstract/Sec. 3.5/Sec. 4 to avoid "consistent across egg sources" claims unless they hold after this fix.

5. **Metric extraction and definitions are not documented at a level that supports verification, and there is a key outcome-definition tension: the paper cannot map "% live birth per transfer" (Secs. 3.1, 3.5), yet Donor_Egg_LB_Rate is later described as "percentage of donor-egg embryo transfer cycles leading to live births," which sounds transfer-denominator-based and risks being confused with the excluded "per transfer" metric (Secs. 3.1–3.2). This raises concern about denominator correctness for multiple metrics (intended vs actual retrieval; transfer-based outcomes) and threatens interpretability of results.**

   Recommendation: Add a mapping table (Sec. 2.2–2.3 or Appendix) listing each analytic metric with: exact NASS Topic/Question/Type/Filter/Breakout fields used; numerator and denominator in words; and any transformations. Explicitly define "intended retrieval" vs "actual retrieval" using NASS documentation and confirm which Cycle_Count corresponds to each metric's denominator. For Donor_Egg_LB_Rate, state the precise numerator/denominator and explain how it differs from (or relates to) the unmappable "% live birth per transfer." Briefly document the search/matching logic that failed for "% live birth per transfer," so readers can reproduce and assess whether it might exist under an alternate label.

6. **Variability estimation is statistically fragile with only 2–3 annual observations per clinic–stratum and with frequent low means/zeros; CV in particular can explode for near-zero means and is undefined when the mean is zero (Secs. 2.4, 3.3, 3.5). The manuscript acknowledges sensitivity but does not quantify instability, specify handling rules for zero means, or show whether key associations are robust to excluding low-mean/low-N strata.**

   Recommendation: In Sec. 2.4, explicitly state: (i) whether SD uses ddof=0 or ddof=1; (ii) how CV is computed when the mean is 0 (exclude vs set missing vs add epsilon), and how "near-zero" is handled; and (iii) how many clinic–strata are affected. In Sec. 3.3 (and/or Appendix), add robustness checks: (a) exclude clinic–strata with mean success rate below a threshold (e.g., $< 1\%$ or $< 2\%$) and re-estimate key correlations/OLS; (b) exclude strata with only 2 years of data; (c) consider alternative dispersion measures (e.g., MAD on annual rates; SD/CV after logit transform of proportions with appropriate continuity correction) and show whether qualitative conclusions persist.

7. **Inferential modeling choices are misaligned with outcome distributions and diagnostics. CV/SD are nonnegative and right-skewed; diagnostic plots show heteroscedasticity/non-normal residuals and likely influential out-**

liers, yet OLS p-values are still used as primary evidence (Secs. 2.5.4, 3.4.3; Figs. 18–21). The state analysis uses many fixed effects with potentially small per-state sample sizes, and the paper runs many tests (metrics × strata × outcomes) with no clear multiple-testing plan (Secs. 2.5.3–2.5.4, 3.4.2–3.4.3).

Recommendation: Either (i) upgrade the modeling strategy or (ii) downgrade inferential claims. Preferably: in Sec. 2.5.4, use heteroscedasticity-robust SEs (e.g., HC3) at minimum; consider modeling $\log(\text{CV} + \epsilon)$ / $\log(\text{SD} + \epsilon)$, Gamma GLM with log link, or robust regression. For geography, consider partial pooling (mixed effects for state) rather than dozens of dummies, or restrict to a smaller pre-specified set of state comparisons. In Sec. 3.4.2–3.4.3, adopt and report a multiple-testing strategy (e.g., FDR within test families) and emphasize effect sizes and uncertainty over isolated $p < 0.05$ findings.

8. **The central "higher volume → higher success-rate variability" result is intriguing but currently under-explained and may reflect mechanical/statistical artifacts (dependence of SD on event counts; CV dependence on mean; regression-to-the-mean; changes in patient mix within stratum; reporting/rounding practices) rather than real instability (Secs. 3.3–3.4.1, 3.5). Without conditioning on mean levels and denominator precision, interpretation remains ambiguous.**

   Recommendation: In Sec. 3.4.1–3.4.3, add analyses that separate level from variability: include mean success rate as a covariate (or stratify by mean-rate bands) when relating volume to SD/CV; test whether volume associations persist within narrower mean ranges. If possible, add simple clinic-level case-mix proxies using NASS (e.g., distribution of age strata across the clinic) to see whether volume–variability associations attenuate. Update Sec. 3.5/Sec. 4 to more explicitly present alternative explanations and avoid causal wording.

9. **Figures and reporting contain multiple consistency/readability issues that reduce actionability and confidence: inconsistent labeling/units (percent vs proportion), mismatches between captions and plotted statistics, p-values shown as 0, conflicting repeated Spearman ρ values (e.g., $\rho = 0.436$ in text vs $\rho = 0.45$ in Fig. 6 caption), missing sample sizes by group, overplotting, and axes compressed by outliers (Figs. 2–3, 6–17; Sec. 3.3–3.4).**

   Recommendation: Systematically audit all figures and captions: standardize terminology (CV, SD, volume), units, and rounding (e.g., report Spearman ρ to 2 d.p. everywhere). Never display $p = 0$; use $p < 1e^{-k}$. Add N per panel/quartile/state in captions or directly on plots. Improve readability (vector/300dpi export, larger fonts, transparency/jitter, axis breaks/insets for extreme outliers). Ensure diagnostic figures (Figs. 18–21) are tied to concrete modeling changes (robust SEs/transformations) rather than presented as a stand-alone caveat.

Minor issues

1. Inclusion/exclusion criteria and missingness are not fully quantified. The paper notes variability requires ≥ 2 years and missing values are kept as NaN (Secs. 2.3–2.4), but does not report how many clinic–strata contribute to each metric by stratum, nor whether inclusion is related to volume or state (Secs. 3.1, 3.3). This can induce selection bias, especially if volume is averaged over "years with metric data."

   Recommendation: Add a table in Sec. 3.1 or Sec. 3.3 showing, for each metric × stratum: number of clinic–strata with 1/2/3 years available; number excluded; and summary of volume for included vs excluded. Briefly discuss implications in Sec. 3.5.

2. State-level analysis presentation is hard to assess because per-state sample sizes and which metric–stratum combinations show significant differences are not summarized, and Kruskal–Wallis is described as comparing "medians" rather than rank distributions (Secs. 2.5.3, 3.4.2).

   Recommendation: In Sec. 3.4.2 (or Appendix), report per-state Ns (or number of states meeting minimum N), list which tests are significant after any correction, and (if doing post hoc comparisons) report which state pairs differ. Rephrase Kruskal–Wallis interpretation as testing distributional/rank differences unless a median-specific post hoc summary is provided.

3. Descriptive results rely heavily on figures and qualitative language ("substantial," "greater"), with few tabular summaries of variability magnitudes (Secs. 3.2–3.4.1).

   Recommendation: Add compact tables in Sec. 3.3–3.4.1 with median (IQR) of CV and SD by metric and stratum (and optionally by volume quartile). Use these to anchor the narrative and facilitate cross-stratum comparison.

4. "Efficiency metrics" are used as a concept but not explicitly defined relative to success metrics; interpretation of variability in these proxies (transfers per intended retrieval; intended retrievals per live birth) could be misunderstood as economic efficiency (Secs. 1, 2.2, 3.2–3.4).

   Recommendation: At first use (Sec. 1 and Sec. 2.2), define "efficiency" as a resource-use proxy rather than cost/technical efficiency. Add 1–2 sentences in Sec. 3.5 explaining how variability in these measures should be interpreted clinically.

5. The omission of % live birth per transfer is likely consequential for interpretation because it is a common patient-facing success metric and helps separate retrieval-stage vs transfer-stage performance; the manuscript mentions the limitation but does not explore implications in depth (Secs. 3.1, 3.5, 4).

   Recommendation: Expand Sec. 3.5 and Sec. 4: explain what additional insight per-transfer variability would provide, and how its absence may bias conclusions. If the metric exists in older years but not 2020–2022 under the same coding, explicitly state

that and why.

6. Rounding/discretization patterns in plots (e.g., horizontal bands) may reflect reporting precision (integer percent, one decimal) or many zeros, which can affect CV/SD and correlation patterns (noted visually in several scatterplots, Sec. 3.3–3.4).

   Recommendation: State the reporting precision of NASS percentages (integer vs decimal) and, if discretization is present, note its implications for dispersion metrics and correlation strength. Consider plotting with jitter/alpha and/or using methods robust to ties in ranks.

7. Reproducibility is described at a high level (Sec. 2.6) but key implementation details (exact extraction filters, software versions) and availability of code are unclear.

   Recommendation: Strengthen Sec. 2.6: specify whether code (and derived nonrestricted data) will be shared; provide repository link if possible; document environment (package versions) and any deterministic seeds; and include pseudocode or scripts for the metric extraction pipeline.

Very minor issues

1. Minor formatting/style issues: split words/line breaks (e.g., "sup\nply"), inconsistent quote styles around variable names, HTML entities (e.g., "< **35**"), and inconsistent heading formatting (e.g., stray "#" in section titles) (Secs. 1–3.4).

   Recommendation: Proofread and standardize typography: fix split words, use consistent formatting for variable names, replace HTML entities with standard symbols ("< **35**", "> **40**"), and harmonize heading styles throughout.

2. Figure ordering and redundancy: some figures appear out of sequence relative to first mention; captions sometimes repeat text and may not match plotted content exactly (Secs. 3.2–3.4.1).

   Recommendation: Reorder figures to match the narrative flow, tighten captions to focus on what is plotted (including N, metric definition, and test used), and verify all in-text references point to the correct figure.

3. Keyword/content mismatch: the keyword list includes "F test" though F-tests are not clearly discussed as part of the analysis narrative (Abstract; Secs. 2.5–3.4).

   Recommendation: Remove "F test" from keywords unless you explicitly report overall model F-statistics in Sec. 3.4.3 and explain their role.

4. A truncated sentence appears in Sec. 3.3 ("mean CV (22.96…") which interrupts interpretation.

   Recommendation: Fix the truncation and ensure all numerical summaries include units (% or percentage points) and complete punctuation.

5. The SD convention is not specified (sample vs population SD), which matters with only 2–3 observations (Sec. 2.4).

   Recommendation: State whether SD uses ddof=0 or ddof=1 and confirm consistency across all metrics/strata.

Key statements and references

- • **The National ART Surveillance System (NASS), maintained by the Centers for Disease Control and Prevention (CDC), collects data on nearly all Assisted Reproductive Technology (ART) cycles performed in the United States, and provides publicly available clinic-level aggregate data suitable for analyses such as this study's 2020–2022 variability assessment.**
- *Reference(s):* CDC
- • **Spearman's rank correlation analysis showed that, for own-egg cycles across all age groups, higher average clinic volume was significantly positively associated with greater year-to-year variability in success rates (e.g., Spearman $\rho = 0.436$ for CV of percentage live birth per intended retrieval in patients < 35 years and $\rho = 0.600$ for SD of percentage live birth per actual retrieval in patients > 40 years, both $p < 0.001$).**
- *Reference(s):* Figure 6, Figure 7, Figure 11
- • **In contrast to success-rate variability, Spearman correlations indicated that higher average clinic volume was significantly negatively associated with variability in certain efficiency metrics for own-egg cycles, such as the coefficient of variation of average transfers per intended retrieval (e.g., $\rho = -0.268$ for patients < 35 years, $p < 0.001$) and the coefficient of variation of average intended retrievals per live birth in younger age groups.**
- *Reference(s):* Figure 1, Figure 5, Figure 10
- • **For donor-egg cycles, higher average clinic volume was also significantly positively associated with greater year-to-year variability in donor-egg live birth rates, with Spearman $\rho = 0.418$ for the coefficient of variation and $\rho = 0.429$ for the standard deviation of the donor-egg live birth rate (both $p < 0.001$), although these estimates are affected by a 2022 data anomaly in NASS where all donor-egg live birth rates were recorded as 0.0%.**
- *Reference(s):* Figure 16, Figure 17
- • **Ordinary Least Squares regression models that included average clinic volume and state as predictors of variability metrics generally had low to moderate explanatory power (R^2 typically < 0.10 to ~ 0.19), but consistently showed that higher average clinic volume had a statistically significant positive coefficient for success-rate variability and a statistically significant**

negative coefficient for variability in some efficiency metrics (e.g., coefficient 0.0041, $p < 0.001$, for SD of percentage live birth per actual retrieval in own-egg cycles age > 40; coefficient -0.0235, $p < 0.001$, for CV of average transfers per intended retrieval in own-egg cycles age 35–37).

- *Reference(s):* Figure 18, Figure 19, Figure 21

Mathematical consistency audit

This section audits **symbolic/analytic** mathematical consistency (algebra, derivations, dimensional/unit checks, definition consistency).

Maths relevance: light

The paper is primarily an applied statistical analysis with limited explicit mathematics: it defines variability measures (CV and SD), describes nonparametric associations (Spearman, Kruskal–Wallis), and outlines OLS regression with a continuous predictor (average clinic volume) and categorical state effects. The main internal-consistency issues are definitional: clinic volume is described in conflicting ways (max vs mean; conditioning on metric availability vs fixed years), and CV computation lacks an explicit rule for zero-mean cases, which are plausible given acknowledged zero inflation and the 2022 donor-egg anomaly.

Checked items

1. ✔ **Coefficient of Variation (CV) definition** (Sec. 2.4, p.3)
    - **Claim:** CV is computed as (Standard Deviation / Mean) $\times 100$ to express relative variability as a percentage of the mean.
    - **Checks:** algebra/definition, unit/dimensional consistency
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** Metric values are numeric on a common scale within each metric (percent kept on 0–100 scale)., Mean is nonzero (or otherwise handled).
    - **Notes:** The formula is correct and dimensionless; with percent-type quantities on 0–100 scale, SD is in percentage points and dividing by the mean yields a unitless ratio, then multiplied by 100 to present percent.
2. ⚠ **Standard deviation (SD) across years definition** (Sec. 2.4, p.3)
    - **Claim:** SD is computed as the standard deviation of a clinic-stratum's metric values across available years (2020–2022).
    - **Checks:** definition completeness, consistency across strata with 2 vs 3 years
    - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
    - **Assumptions/inputs:** At least two yearly observations exist for that clinic-stratum-metric., A specific SD convention (sample vs population) is used consistently.

- **Notes:** The SD concept is fine, but the paper does not specify whether SD uses ddof=0 or ddof=1. With only **2–3** data points, this materially changes SD and therefore CV; the analytic definition is incomplete.

3. ✔ **Eligibility rule for variability calculation ($\geq$ 2 years)** (Sec. 2.4, p.3)
    - **Claim:** Variability (CV, SD) is computed only when at least two of the three years have reported data for that clinic-stratum-metric; otherwise excluded.
    - **Checks:** logical consistency, handling of missingness (conceptual)
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** Missing values indicate absent reporting and are excluded from the across-year computation., At least two remaining values allow SD/CV computation.
    - **Notes:** The rule is logically consistent with computing dispersion measures across time and avoids undefined SD for $n < 2$.

4. ✔ **Percent scale normalization for metrics** (Sec. 2.3, p.3)
    - **Claim:** Percentages are represented uniformly on a **0–100** scale before computing variability.
    - **Checks:** unit/scale consistency
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** All percent-type metrics are consistently converted to the same scale., Non-percent metrics (e.g., averages per cycle) remain on their natural scale.
    - **Notes:** Keeping percent outcomes on a common **0–100** scale makes SD interpretable as percentage points. (Numerical correctness of conversions cannot be checked here.)

5. ✔ **Definition of Avg_Clinic_Volume in Methods** (Sec. 2.4, p.3)
    - **Claim:** Avg_Clinic_Volume is the mean of reported Cycle_Count values for a clinic-stratum across the years for which metric data was available.
    - **Checks:** definition clarity, cross-section consistency
    - **Verdict:** PASS; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Cycle_Count is available per clinic-year-stratum., The set of years used in the average may differ by metric depending on missingness.
    - **Notes:** This definition is coherent on its own, but it implies Avg_Clinic_Volume is metric-dependent (because 'years for which metric data was available' depends on the metric), which should be stated explicitly if used that way.

6. ✖ **Clinic volume definition inconsistency (max vs mean; averaging set)** (Sec. 3.1, p.4; Sec. 3.4, p.6; contrasted with Sec. 2.4, p.3)
    - **Claim:** Results describe clinic volume as the maximum cycle count for a clinic-year-stratum (Stratum_Cycle_Count), while also using Avg_Clinic_Volume as an average across 2020–2022.
    - **Checks:** definition consistency, symbol/variable consistency
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** Per-year stratum cycle count is uniquely defined; if repeated across multiple metric rows, taking max is equivalent to taking any identical repeated value., Averaging is performed over a stated year set.
    - **Notes:** Methods (mean over years with metric data) conflict with Results (per-year value described as a maximum; average described as across 2020–2022). These are not guaranteed to match and directly affect the main predictor in Spearman/OLS analyses.
7. ⚠ **CV well-definedness under zero/near-zero means** (Sec. 2.4, p.3; anomaly discussed Sec. 3.2 and Sec. 3.5, p.5 and p.11)
    - **Claim:** CV is used broadly across strata despite acknowledged zero-inflation and a 2022 donor metric reported as **0.0%** for all clinics.
    - **Checks:** domain/edge-case analysis, definition completeness
    - **Verdict:** UNCERTAIN; confidence: high; impact: critical
    - **Assumptions/inputs:** CV uses division by the mean across available years., Mean may be zero for some clinic-strata.
    - **Notes:** If the across-year mean equals **0**, CV is undefined; if very small, CV can be arbitrarily large. The paper notes sensitivity but does not specify a computational/analytic rule for these cases, so CV-based analyses are not fully defined.
8. ✔ **Interpretation of SD for percent outcomes as 'percentage points'** (Sec. 3.3, p.6)
    - **Claim:** Median SD values for percent success-rate metrics are interpreted as absolute fluctuations in percentage points.
    - **Checks:** unit/interpretation consistency
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** Success-rate metrics are stored on a **0–100** percent scale (not **0–1**).
    - **Notes:** Given the stated **0–100** scaling, SD is naturally in percentage points; the interpretation matches the units.
9. ✔ **Spearman correlation usage** (Sec. 2.5.2, p.4; results discussion Sec. 3.4.1, p.6–p.8)

- **Claim:** Spearman's rank correlation is used to assess monotone association between variability measures (CV/SD) and Avg_Clinic_Volume.
- **Checks:** method/definition sanity-check
- **Verdict:** PASS; confidence: high; impact: minor
- **Assumptions/inputs:** Observations (clinic-strata) are treated as independent for correlation calculation.
- **Notes:** No algebraic issues; the method choice is consistent with the stated goal of robust association assessment.

10. ⚠ **Kruskal–Wallis described as median comparison** (Sec. 2.5.3, p.4; Sec. 3.4.2, p.8)

    - **Claim:** Kruskal–Wallis compares median variability across groups (states or volume quartiles).
    - **Checks:** conceptual/statistical statement precision
    - **Verdict:** UNCERTAIN; confidence: high; impact: minor
    - **Assumptions/inputs:** Groups are independent; test is applied to rank-based distributions.
    - **Notes:** The statement is imprecise: Kruskal–Wallis is a rank-based test for distributional differences and is not strictly a 'median test' as written. This is primarily wording; the underlying computation may still be appropriate.

11. ✔ **OLS model specification with state dummies and reference state** (Sec. 2.5.4, p.4; Sec. 3.4.3, p.9–p.10)

    - **Claim:** OLS models regress CV or SD on Avg_Clinic_Volume and LocationAbbr treated as categorical via dummies with a reference state.
    - **Checks:** notation/definition sanity-check, model identifiability
    - **Verdict:** PASS; confidence: medium; impact: moderate
    - **Assumptions/inputs:** An intercept is included and one state dummy is omitted (reference) to avoid perfect multicollinearity.
    - **Notes:** The described setup is internally consistent and identifiable if implemented with a reference category. No explicit regression equation is given, so only the specification can be sanity-checked.

12. ✖ **Truncated mean CV sentence (donor metric)** (Sec. 3.3, p.6)

    - **Claim:** The text reports a mean CV value but the sentence is truncated ('mean CV (22.96').
    - **Checks:** notation/typography
    - **Verdict:** FAIL; confidence: high; impact: minor
    - **Notes:** This is a typographical/incomplete statement that obscures the reported quantity's units and exact value formatting. It does not create an algebraic contradiction but reduces interpretability.

Limitations

- The paper contains very few explicit equations/derivation steps beyond definitions; many analytic details (e.g., exact SD convention, exact handling of zero means for CV, exact year sets used for averaging volumes) are not explicitly stated, limiting verifiability.
- No explicit regression equations (design matrix, transformations, intercept handling) are provided; the audit can only assess the verbal specification for internal consistency.
- Numerical values, plots, and reported test statistics/p-values are not checked by design (scope excludes numerical verification).

Numerical results audit

This section audits **numerical/empirical** consistency: reported metrics, experimental design, baseline comparisons, statistical evidence, leakage risks, and reproducibility.

Seventeen numeric checks were identified. One internal-consistency check failed (a repeated Spearman ρ value differs between text and figure caption beyond plausible rounding). The remaining sixteen items could not be recomputed and are therefore uncertain given the available inputs.

Checked items

1. ⚠ **CAND-001** (Page 4 (Section 3.1 Data cohort and preparation))
    - **Claim:** "The initial dataset comprised 1,126,080 records."
    - **Checks:** record_count_exact_match
    - **Verdict:** UNCERTAIN
    - **Notes:** Cannot recompute: required dataset/CSV not available (script is constrained to use only PAYLOAD and cannot read files).
2. ⚠ **CAND-002** (Page 4 (Section 3.1 Data cohort and preparation))
    - **Claim:** "A total of 510 unique clinics were identified across the U.S. states and territories."
    - **Checks:** unique_count_recompute
    - **Verdict:** UNCERTAIN
    - **Notes:** Cannot recompute: required dataset/CSV not available (script is constrained to use only PAYLOAD and cannot read files).
3. ⚠ **CAND-003** (Page 4 (Section 3.1 Data cohort and preparation))
    - **Claim:** "The data were reshaped into a wide format, with each row representing a unique clinic-year-stratum combination, totaling 6,800 such combinations."
    - **Checks:** unique_combination_count_recompute

- **Verdict:** UNCERTAIN
- **Notes:** Cannot recompute: required dataset/CSV not available (script is constrained to use only PAYLOAD and cannot read files).

4. ⚠ **CAND-004** (Page 5 (Section 3.2 Descriptive analysis of ART metrics), example for 2020 own eggs)
    - **Claim:** "In 2020, the mean Perc_LB_IntendedRetrieval ranged from 15.9% for the $<$ **35** age group to 1.3% for the $>$ **40** age group."
    - **Checks:** range_consistency_and_group_means
    - **Verdict:** UNCERTAIN
    - **Notes:** Cannot recompute: required dataset/CSV not available (script is constrained to use only PAYLOAD and cannot read files).
5. ⚠ **CAND-005** (Page 5 (Section 3.2 Descriptive analysis of ART metrics), 2022 shift for age 35-37)
    - **Claim:** "...a marked increase in mean Perc_LB_IntendedRetrieval for the 35-37 group (27.2%) compared to previous years..."
    - **Checks:** group_mean_recompute_and_year_comparison
    - **Verdict:** UNCERTAIN
    - **Notes:** Cannot recompute: required dataset/CSV not available (script is constrained to use only PAYLOAD and cannot read files).
6. ⚠ **CAND-006** (Page 5 (Section 3.2 Descriptive analysis), donor egg rates 2020/2021)
    - **Claim:** "For donor egg cycles, the mean Donor_Egg_LB_Rate was 3.2% in 2020 and 3.9% in 2021 across reporting clinics."
    - **Checks:** group_mean_recompute
    - **Verdict:** UNCERTAIN
    - **Notes:** Cannot recompute: required dataset/CSV not available (script is constrained to use only PAYLOAD and cannot read files).
7. ⚠ **CAND-007** (Page 5 (Section 3.2 Descriptive analysis), donor anomaly and count)
    - **Claim:** "...in 2022, where the Donor_Egg_LB_Rate was reported as 0.0% for all 457 clinic-year-stratum instances available in the descriptive analysis."
    - **Checks:** all_values_equal_and_count_recompute
    - **Verdict:** UNCERTAIN
    - **Notes:** Cannot recompute: required dataset/CSV not available (script is constrained to use only PAYLOAD and cannot read files).
8. ⚠ **CAND-008** (Page 6 (Section 3.3 Variability), own-egg Perc_LB_IntendedRetrieval variability medians)
    - **Claim:** "For own-egg success rates (Perc_LB_IntendedRetrieval), the median SD across clinic-strata was 3.70... The median CV was 86.6%..."

- **Checks:** median_recompute_from_defined_formula
- **Verdict:** UNCERTAIN
- **Notes:** Cannot recompute: required dataset/CSV not available (script is constrained to use only PAYLOAD and cannot read files).

9. ⚠ **CAND-009** (Page 6 (Section 3.3 Variability), own-egg Perc_LB_ActualRetrieval variability medians)
    - **Claim:** "Similarly, Perc_LB_ActualRetrieval showed a median SD of 3.56 and a median CV of 25.1%."
    - **Checks:** median_recompute_from_defined_formula
    - **Verdict:** UNCERTAIN
    - **Notes:** Cannot recompute: required dataset/CSV not available (script is constrained to use only PAYLOAD and cannot read files).
10. ⚠ **CAND-010** (Page 6 (Section 3.3 Variability), Avg_Transfers_IntendedRetrieval variability medians)
    - **Claim:** "Avg_Transfers_IntendedRetrieval had a median SD of 0.25 and CV of 53.9%..."
    - **Checks:** median_recompute_from_defined_formula
    - **Verdict:** UNCERTAIN
    - **Notes:** Cannot recompute: required dataset/CSV not available (script is constrained to use only PAYLOAD and cannot read files).
11. ⚠ **CAND-011** (Page 6 (Section 3.3 Variability), Avg_IntendedRetrievals_LB variability medians)
    - **Claim:** "Avg_IntendedRetrievals_LB showed a median SD of 1.62 and a high median CV of 92.9%."
    - **Checks:** median_recompute_from_defined_formula
    - **Verdict:** UNCERTAIN
    - **Notes:** Cannot recompute: required dataset/CSV not available (script is constrained to use only PAYLOAD and cannot read files).
12. ⚠ **CAND-012** (Page 6 (Section 3.3 Variability), donor egg variability means vs anomaly)
    - **Claim:** "For the Donor_Egg_LB_Rate, the median SD and CV were reported as 0.0%... However, the mean SD (3.70%) and mean CV (22.96..."
    - **Checks:** mean_and_median_recompute_from_defined_formula
    - **Verdict:** UNCERTAIN
    - **Notes:** Cannot recompute: required dataset/CSV not available (script is constrained to use only PAYLOAD and cannot read files).

13. ✖ **CAND-013** (Page 6-7 (Section 3.4.1 Association with clinic volume; Figures 6-8 text))
    - **Claim:** Spearman Rho values are reported inconsistently across text/figure captions for similar relationships (e.g., CV_Perc_LB_ActualRetrieval for age < 35 reported as 0.436 in text vs 0.45 in Figure 6 caption; age 38–40 reported as 0.49 in Figure 7 caption).
    - **Checks:** internal_consistency_of_repeated_statistic
    - **Verdict:** FAIL
    - **Notes:** Discrepancy exceeds stated plausible rounding/tolerance for repeated statistic.
14. ⚠ **CAND-014** (Page 6 (Section 3.4.1 negative correlation example))
    - **Claim:** "CV_Avg_Transfers_IntendedRetrieval correlated negatively with volume... (e.g., Rho=-0.268 for < 35)."
    - **Checks:** spearman_correlation_recompute
    - **Verdict:** UNCERTAIN
    - **Notes:** Cannot recompute: required dataset/CSV not available (script is constrained to use only PAYLOAD and cannot read files).
15. ⚠ **CAND-015** (Page 10 (Section 3.4.3 OLS example coefficient))
    - **Claim:** "...in the model for SD_Perc_LB_ActualRetrieval (Own, > 40), the coefficient for Avg_Clinic_Volume was 0.0041 ($p < 0.001$)."
    - **Checks:** regression_coefficient_recompute
    - **Verdict:** UNCERTAIN
    - **Notes:** Cannot recompute: required dataset/CSV not available (script is constrained to use only PAYLOAD and cannot read files).
16. ⚠ **CAND-016** (Page 10 (Section 3.4.3 OLS example coefficient))
    - **Claim:** "...-0.0235 for CV_Avg_Transfers_IntendedRetrieval (Own, 35–37), $p < 0.001$..."
    - **Checks:** regression_coefficient_recompute
    - **Verdict:** UNCERTAIN
    - **Notes:** Cannot recompute: required dataset/CSV not available (script is constrained to use only PAYLOAD and cannot read files).
17. ⚠ **CAND-017** (Page 9 (Section 3.4.3 Multivariable OLS regression analysis))
    - **Claim:** "R-squared values, typically ranging from below 0.10 to around 0.19."
    - **Checks:** range_check_over_reported_models
    - **Verdict:** UNCERTAIN

- **Notes:** Cannot recompute: required dataset/CSV not available (script is constrained to use only PAYLOAD and cannot read files).

Limitations

- Only parsed PDF text was available; no access to the underlying NASS CSV or the authors' code, so all recomputation checks are contingent on having those files.
- The PDF provides many results without full tabular outputs (no tables of descriptive stats, CV/SD summaries, or regression outputs), limiting the number of internal arithmetic cross-checks possible purely within-document.
- Some metric names/stratum construction rules are described at a high level (e.g., how donor vs own is identified, how cycle count is summarized), so recomputation must mirror the authors' exact mapping choices to match reported numbers.
- Several candidate checks could not be recomputed because the required dataset/CSV was not available to the execution script (constrained to use only PAYLOAD and cannot read files).
- Computational environment/package versions listed in the manuscript cannot be verified from the provided inputs without environment metadata or code.
- Figure-based qualitative distribution claims and exact plot-implied values cannot be verified without extracting plot data or pixels.
- Some definitional statements (e.g., clinic volume as a max cycle count per clinic-year-stratum) and exact p-values (e.g., Kruskal-Wallis $P = 0.0$) cannot be confirmed without the underlying dataset and precise implementation details.