Skepthical review: *Efficiency Analysis of US ART Clinics: A Data Envelopment Analysis Approach (2020-2022)*

Summary

The manuscript applies an input-oriented BCC (VRS) Data Envelopment Analysis (DEA) model to benchmark U.S. Assisted Reproductive Technology (ART) clinics using CDC NASS data (2020–2022). Decision-making units (DMUs) are defined at the clinic–year–age-group level ($<$ **35**, **35–37**, **38–40**, $>$ **40**), with one input (intended own-egg retrieval cycles) and one output (live births), where live births are constructed from the reported live-birth percentage multiplied by the cycle count (Sec. 2.1–2.3). The paper reports strongly right-skewed distributions, many zero/near-zero outcomes (especially at older ages), low mean/median efficiency, a small fraction of frontier units, and declining efficiency with age (Sec. 3.2–3.4), plus a sensitivity analysis highlighting very low efficiency among DMUs with zero observed live births (Sec. 3.5). The topic is policy-relevant and the dataset is valuable, but as currently specified the 1-input/1-output DEA largely risks collapsing to a transformed success-rate ranking (and, in practice, to artifacts of output construction/rounding and suppression), limiting interpretability as "technical efficiency" in the operations/economics sense. Strengthening measurement (especially output construction), clarifying what DEA adds beyond live-birth rates, improving risk adjustment/interpretation, adding robustness/inference, and cleaning up presentation/reproducibility would substantially improve credibility and impact (Sec. 1–4).

Strengths

- Uses a large, recent national dataset (NASS 2020–2022) with clear stratification by clinically salient age groups and by year (Sec. 2.1–2.2).
- Implements a standard DEA linear program (input-oriented BCC/VRS) and provides a high-level description of the computational approach using `scipy.optimize.linprog` and parallelization (Sec. 2.3–2.4).
- Descriptive statistics and visualizations clearly communicate the heavy right-skewness and zero-inflation in outcomes, which are central to interpreting any efficiency analysis in this setting (Sec. 3.2).
- The paper explicitly investigates the role of zero live-birth observations via a sensitivity analysis, which is an important diagnostic given the data structure (Sec. 3.5).
- Internal dataset-size accounting and several aggregate numerical checks appear consistent (e.g., DMU counts and reported efficient shares), which increases confidence that the pipeline runs as described.

Major issues

1. **With one input and one output, the input-oriented VRS (BCC) DEA is often very close to a (piecewise) transformation of the output/input ratio —here essentially the live-birth rate—so the analysis may be rediscovering the distribution of live-birth rates (and its zero mass) rather than identifying multi-factor technical efficiency (Sec. 1, Sec. 2.3, Sec. 3.3–3.4). Because the output is constructed as (live-birth rate × cycle count), the model also mechanically ties output to input, further compressing what DEA can learn beyond "which clinics have the best reported rate within a stratum."**

   Recommendation: In Sec. 1 and Sec. 4.3, explicitly discuss what DEA adds in a 1×1 setting and why DEA (rather than direct rate benchmarking) is the preferred tool here. Add an empirical diagnostic in Sec. 3.3 (or a new Sec. 3.7): within each year× age stratum, plot/correlate DEA efficiency vs. live-birth rate (live births per intended retrieval) and explain deviations (if any). If the intended contribution is true efficiency benchmarking, add a robustness/extension with additional outputs and/or inputs available in NASS (or clearly state they are unavailable), so DEA is substantively necessary rather than a reparameterized rate ranking (Sec. 2.1–2.3).

2. **Output construction has a critical unit/rounding problem that can materially distort efficiency scores, especially for low-volume clinics and older age groups (Sec. 2.2, Sec. 3.2.3–3.2.4, Sec. 3.5). The manuscript defines live births as ("% live births per intended retrieval") × (cycle count) without clearly dividing by 100, and the unstructured report indicates the result is then rounded to an integer. This can create many artificial zeros whenever expected births are < 0.5 and compound discretization error if the published percentage is itself rounded. The strong results about 'zero live births' (Sec. 3.5) may therefore be driven partly by construction artifacts rather than clinical performance.**

   Recommendation: In Sec. 2.2, define the output precisely and fix units: \text{Output_LiveBirths} = (\text{LBRate_percent}/100) \times \text{Cycle_Count} if LBRate is a percent. Do not round outputs; DEA permits nonnegative real outputs. Re-run the full analysis with continuous outputs and report how (i) the fraction of zero-output DMUs changes, (ii) efficiency distributions change, and (iii) the Sec. 3.5 sensitivity conclusions change. As an explicit robustness check (Appendix or Sec. 3.7), compare efficiency under (a) continuous outputs, (b) current integer-rounded outputs, and (c) alternative rounding rules (floor/ceiling) to quantify sensitivity.

3. **Interpretation of input-oriented efficiency is not well aligned with ART clinical objectives, and the chosen orientation/returns-to-scale assumptions are under-justified (Sec. 1, Sec. 2.3, Sec. 4.3). Input orientation asks how much intended retrieval cycles could be reduced holding live births fixed—**

yet cycles are patient treatments/demand-driven rather than an easily "minimized" resource, and many readers will interpret the results as 'quality' rather than 'input contraction.'

Recommendation: Expand justification in Sec. 2.3 for (i) input orientation and (ii) VRS (BCC) with citations to health-care DEA practice. Add at least one robustness model that better matches the clinical aim of improving success per attempt: e.g., output-oriented BCC (maximize live births given cycles), and optionally CRS vs VRS. Summarize whether the core patterns (low scores, age gradient, frontier shares) persist (Sec. 3.7 / Appendix). Tighten language in Sec. 3.3–3.6 and Sec. 4.2–4.3 so θ is interpreted as radial input reduction (not "ability to increase outputs") unless an output-oriented model is also presented.

4. **Case-mix adjustment beyond coarse age bands is insufficient, so 'inefficiency' may largely reflect patient severity/selection and clinic environment rather than performance (Sec. 1, Sec. 2.1–2.2, Sec. 3.3–3.6, Sec. 4.3). Within age groups, outcomes vary by diagnosis, ovarian reserve, prior ART history, use of ICSI/PGT, embryo transfer practices, comorbidities, and socioeconomic factors; ignoring these risks confounding and can create ethically problematic incentives if interpreted as rankings.**

   Recommendation: Create a dedicated Limitations subsection in Sec. 4.3 explicitly stating that scores are conditional on minimal risk adjustment and should not be interpreted as causal performance. If additional variables are available in NASS (or via linkage), add either: (i) a two-stage analysis (DEA then regression of scores on environmental/case-mix proxies), (ii) a conditional/non-discretionary-input DEA variant, or (iii) restrictions to more homogeneous subgroups. At minimum, add exploratory stratifications/correlates in Sec. 3.6 (e.g., by clinic volume, region, ownership if available) and qualify all cross-clinic comparisons accordingly.

5. **The DMU definition (clinic–year–age group) implies stratum-specific frontiers and complicates cross-stratum statements (e.g., 'efficiency decreases with age') because these are comparisons across separate DEA runs rather than a single unified technology (Sec. 2.1, Sec. 3.3–3.4). Additionally, VRS 'scale' in this setup is essentially the number of intended retrievals within the stratum, which may not map cleanly to clinic operational scale.**

   Recommendation: In Sec. 2.1 and at the start of Sec. 3.3, clarify that efficiencies are computed relative to a year×age-specific frontier and are not directly comparable across strata unless you adopt pooling, a meta-frontier, or a normalization strategy. If the paper's narrative emphasizes age gradients, add a robustness check using pooled models (e.g., include age group as a categorical environmental factor, or estimate a meta-frontier) and explicitly discuss what changes. Consider adding a clinic-level aggregation robustness run to show how results differ when the DMU is 'clinic-year' (Sec. 3.7 / Appendix).

6. **Zero and near-zero outputs are prevalent and materially shape the frontier, but the manuscript does not fully explain the mathematical behavior of the BCC model with $y_o = 0$, nor separate true zeros from construction/suppression-induced zeros (Sec. 2.2–2.3, Sec. 3.2.3–3.2.4, Sec. 3.5). Some statements risk implying that $y_o = 0$ forces very low θ, which is not mathematically necessary under BCC-I.**

   Recommendation: In Sec. 2.3 and Sec. 3.5, add a short analytic explanation of how the LP behaves when $y_o = 0$ (output constraint becomes nonbinding; efficiency depends on input minimality within the convex hull). In Sec. 3.2.3–3.2.4, report the proportion of DMUs with (constructed) zero output by age and year and—after fixing output construction (continuous, no rounding)—reassess how many zeros remain. Add robustness checks excluding (a) zero-output DMUs and (b) very small \text{Cycle_Count} DMUs (where discretization dominates), and report how frontier composition and mean/median θ change (Sec. 3.7 / Appendix).

7. **Suppressed/missing NASS cells ('$*$', '$-$') are dropped, which likely removes small-volume clinics/strata non-randomly and can bias efficiency distributions and frontier identification (Sec. 2.2, Sec. 3.1–3.2). Because suppression is often related to privacy thresholds, the missingness mechanism is plausibly informative.**

   Recommendation: In Sec. 2.2 and Sec. 3.1–3.2, quantify suppression and deletions by year and age group, and compare \text{Cycle_Count} distributions for kept vs dropped records to assess selection. Consider sensitivity bounds or interval imputation approaches consistent with suppression rules (even a simple 'best/worst case' for suppressed cells), or clearly state the likely direction of bias (e.g., under-representing low-volume clinics).

8. **No statistical inference or stability analysis is provided, despite DEA's sensitivity to sampling variation, measurement error, and outliers—especially with many small-volume DMUs and constructed outputs (Sec. 3.3–3.4). Statements that year-to-year changes are 'minor' are not supported by formal uncertainty quantification (Sec. 3.4.1).**

   Recommendation: Add a robustness/inference component: within each stratum, use bootstrap DEA (e.g., Simar–Wilson style) or at minimum resampling-based confidence intervals for mean/median efficiency and bias-corrected scores. Complement with stability checks: trimming/winsorizing extreme DMUs, excluding very small \text{Cycle_Count}, comparing DEA vs FDH. In Sec. 3.4.1, either provide uncertainty intervals for year comparisons or explicitly label the temporal analysis as purely descriptive and avoid inferential wording.

9. **Positioning, novelty, and implications are underdeveloped, and the manuscript lacks substantive engagement with prior DEA-in-health/ART efficiency literature; practical meaning of $\theta \approx 0.25$ is not concretely interpreted (Sec. 1–2, Sec. 3.6, Sec. 4.2–4.3). This also heightens the risk that readers treat the results as clinic 'rankings' rather than conditional benchmarks with major limitations.**

   Recommendation: Add a Related Work subsection (Sec. 1.1 or Sec. 2.x) summarizing DEA applications in health care and any ART/fertility clinic benchmarking, including typical input/output choices and risk-adjustment practices. Near the end of Sec. 1, state clear research questions and contributions. In Sec. 3.6 and Sec. 4.3, translate efficiency scores into concrete DEA interpretations (input contraction under input orientation; output expansion under output orientation if added), provide at least one worked example with peer/reference sets, and add a dedicated Ethical/Policy Considerations paragraph cautioning against simplistic rankings and noting missing safety/equity/patient-centered outcomes.

10. **Reproducibility and readability are impaired by reliance on internal file paths for key results and by missing/unclear data/code availability information (Sec. 3.1–3.5, Sec. 2.4).**

    Recommendation: Replace internal path references (e.g., `data/dea_analysis_results/...`) with numbered tables/figures in the paper or appendices, and include key numeric summaries directly in Sec. 3 (e.g., quartiles of θ, zero-output shares, frontier counts). Add a Data and Code Availability statement (end of Sec. 2 or in Sec. 4) describing what can be shared, with a public repository link and enough documentation to rerun the pipeline (including versions, solver method/options, and data-processing scripts).

Minor issues

1. Several manuscript metadata elements appear to be placeholders or incorrect, including irrelevant keywords ("Cosmology, Orbits, Relativity…") and an affiliation line that reads as non-academic placeholder text (Abstract/front matter).

   Recommendation: Replace keywords with ART/DEA-relevant terms and correct affiliations/author information to match journal expectations before submission.

2. Figures are often overcrowded or too small (e.g., Figures 1, 3–10), with illegible labels and missing sample-size annotations; inconsistent axis scaling/ordering may mislead cross-panel comparisons (Sec. 3.2–3.5).

   Recommendation: Split multi-panel figures, increase font sizes/export resolution ($\geq$ **300** dpi or vector), standardize age-group ordering and axis limits where comparisons are intended, and add 'n' annotations per facet. Ensure captions state binning/normalization choices and define key metrics.

3. Data preprocessing details are not sufficiently explicit for replication (suppressed values handling, exact NASS filters, clinic identifiers, and exclusions), and the paper does not fully characterize excluded records (Sec. 2.1–2.2, Sec. 3.1–3.2).

   Recommendation: In Sec. 2.2, provide step-by-step filtering rules with exact variable names/values, explicit conversion rules for '∗'/'−', handling of clinic identifiers/locations, and a flow table of exclusions by year×age group in Sec. 3.1.

4. The operational definition of "efficient" DMUs (e.g., $\theta \geq 0.9999$) is not introduced where efficiency results are first reported, and numerical tolerance is not justified (Sec. 2.3, Sec. 3.3).

   Recommendation: Define the efficiency threshold at the start of Sec. 3.3, justify the tolerance based on solver precision, and ensure all efficient counts/percentages consistently use that rule.

5. The DEA formulation is presented redundantly in multiple places, which interrupts narrative flow (Sec. 1 vs Sec. 2.3).

   Recommendation: Keep a brief conceptual description in Sec. 1 and move the full LP specification to Sec. 2.3, with a clear cross-reference.

6. Solver and computation details are incomplete (e.g., which `linprog` method, tolerances, infeasibility handling, runtime), limiting reproducibility for readers attempting to scale the approach (Sec. 2.4).

   Recommendation: In Sec. 2.4 (or Appendix), state the `linprog` method/options, tolerances, how infeasible/unbounded cases are handled (and whether any occurred), and report approximate runtimes/hardware. Consider adding short pseudocode for the DEA loop and parallelization.

7. Definitions of ART outcome measures are not fully clarified (e.g., whether "live births" correspond to deliveries vs infants; handling of multiple births), which matters for interpretation (Sec. 2.1–2.2).

   Recommendation: Add a brief definitions paragraph citing NASS documentation specifying precisely what the reported live-birth metric counts and how it is aggregated.

Very minor issues

1. Formatting/typographical inconsistencies persist (mixed quoting/backticks around column names, broken line breaks within words, truncated file names, inconsistent LaTeX/plain-text for age groups and percentages) (Sec. 2.1–2.4, Sec. 3.2–3.5).

   Recommendation: Proofread and standardize formatting throughout: consistent $\texttt{}` for variables, fix broken words, remove truncated path fragments from the main text, and harmonize age-group and percent formatting.

2. Headings and captions show inconsistent styles (stray `#`, mixed numbering, caption-like lines appearing as headings) (Sec. 3.2–3.5).

   Recommendation: Align headings/captions with the target journal template: consistent numbering, remove stray markers, and ensure each figure caption is concise and self-contained.

3. The optimization problem statement omits explicit bounds/wording commonly included for completeness (e.g., $\theta \geq 0$) and slightly overbroad statements about θ ranging down to 0 under the paper's positive-input filtering (Sec. 2.3).

   Recommendation: Add $\theta \geq 0$ explicitly (or note it is implied) and tighten the discussion of θ's attainable range given the data cleaning and reference-set inclusion.

Key statements and references

- ✔ **Data Envelopment Analysis (DEA) is a non-parametric method that evaluates the efficiency of decision-making units by comparing their input–output relationships to an efficiency frontier representing best observed practices, and unlike traditional statistical methods it does not require prior assumptions about the underlying production function or the distribution of errors, which makes it suitable for complex multi-input, multi-output systems such as ART clinics.**
- *Reference(s):* (none)
- *Justification:* No valid PDFs found; assumed supported.

- ✔ **The National ART Surveillance System (NASS) dataset, maintained by the U.S. Centers for Disease Control and Prevention (CDC), provides comprehensive clinic-level information on Assisted Reproductive Technology procedures performed in the United States for reporting years including 2020–2022, enabling analyses restricted to cycles where patients use their own eggs and stratified by patient age group.**

- *Reference(s):* (none)
- *Justification:* No valid PDFs found; assumed supported.

- ✔ **The well-documented impact of patient age on ART outcomes, with live birth rates per intended retrieval declining systematically as age increases (e.g., mean live birth rate of about 22.46% for patients < 35 years versus about 1.95% for patients > 40 years in 2020), motivates stratifying efficiency analysis by age group to control for age-related effects on ART success rates.**

- *Reference(s):* (none)
- *Justification:* No valid PDFs found; assumed supported.

Mathematical consistency audit

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

Maths relevance: light

The paper's core mathematics is a single-input single-output input-oriented BCC (variable returns to scale) DEA linear program used to compute efficiency scores θ for clinic-year-age-group DMUs. Most of the document is descriptive/statistical reporting; the main analytic risks are definition/units consistency for the constructed output variable and correct interpretation of the DEA measure under special cases (notably zero outputs).

Checked items

1. ✔ **BCC-I DEA primal LP formulation** (Optimization problem shown in Introduction/early text (p.2) and again in Sec. 2.3 (p.3))
    - **Claim:** Clinic o's efficiency θ is obtained by minimizing θ subject to weighted inputs not exceeding θ-scaled input of o, weighted outputs at least the output of o, VRS convexity (sum λs $= 1$), and λ nonnegativity.
    - **Checks:** algebra/LP constraint structure, definition consistency
    - **Verdict:** PASS; confidence: high; impact: critical
    - **Assumptions/inputs:** Single input x and single output y per DMU as stated, Reference set includes all DMUs $j = 1..N$ in the stratum (implicitly including o), Inputs are nonnegative (and later filtered to be strictly positive)
    - **Notes:** The constraints match a standard input-oriented BCC/VRS DEA primal form for a single input/output case, and symbols are defined consistently (x_j, y_j, x_o, y_o, λs, N).
2. ✔ **VRS convexity constraint interpretation** (Sec. 2.3, paragraph explaining the third constraint (p.3))
    - **Claim:** The constraint $\sum_j \lambda_j = 1$ enforces VRS and makes the frontier a convex hull rather than a ray from the origin.
    - **Checks:** conceptual/analytic consistency
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** Nonnegative λs, Technology set constructed as convex combinations of observed DMUs
    - **Notes:** Given $\lambda \geq 0$, the sum-to-1 constraint produces convex combinations (convex hull) rather than a cone through the origin (CRS-like).
3. ✖ **Output construction from percentage rate and cycle count** (Sec. 2.2 (p.2) and Sec. 3.1 bullet defining \text{Output_LiveBirths} (p.4))

- **Claim:** Live births are calculated by multiplying "% Live Births per Intended Retrieval" (\text{Data_Value_num}) by \text{Cycle_Count} and rounding.
- **Checks:** units/dimensional consistency, definition consistency
- **Verdict:** FAIL; confidence: high; impact: critical
- **Assumptions/inputs:** \text{Data_Value_num} is described as a percentage (e.g., **22.46%**), \text{Cycle_Count} is a count of intended retrieval cycles
- **Notes:** A percentage quantity (in % units) must be converted to a unitless proportion before multiplying by a count to yield a count. As written, $\text{Output_LiveBirths} = (\text{percent})\times(\text{count})$ is scale-inconsistent unless \text{Data_Value_num} is already in $[0, 1]$. The paper simultaneously labels the variable as a percent and gives examples in percent form, but never states a $/100$ conversion or that \text{Data_Value_num} is a proportion.

4. ✔ **Rounding live births to nearest integer** (Sec. 2.2 (p.2) and reiterated in Sec. 3.1 (p.4))
   - **Claim:** The computed live births are rounded to the nearest whole number and cast to integer because live births are discrete counts.
   - **Checks:** modeling/analytic compatibility
   - **Verdict:** PASS; confidence: medium; impact: minor
   - **Assumptions/inputs:** Computed live-birth expectation may be non-integer before rounding, DEA LP accepts real-valued outputs
   - **Notes:** Rounding is not an algebraic inconsistency with the DEA LP (which allows real inputs/outputs), but it changes the constructed outputs and can alter the shape of the frontier. The main mathematical risk here is not inconsistency but lack of justification/sensitivity discussion.
5. ✔ **Feasibility of each DMU's LP** (Sec. 2.3 constraint set with $j = 1..N$ (p.3))
   - **Claim:** The DEA LP can be solved for each clinic/DMU within a stratum.
   - **Checks:** feasibility sanity check
   - **Verdict:** PASS; confidence: medium; impact: moderate
   - **Assumptions/inputs:** The evaluated DMU o is included among the N DMUs in the stratum
   - **Notes:** If DMU o is part of the reference set, choosing $\lambda_o = 1$ yields feasibility with $\theta = 1$. The paper implies (but does not explicitly state) that o is included in $j = 1..N$.
6. ✔ **Range/interpretation of efficiency score θ** (Sec. 2.3 explanation of θ (p.3) and Sec. 3.3 (p.6))

- **Claim:** θ ranges from 0 to 1; $\theta = 1$ indicates efficiency; $\theta < 1$ indicates inefficiency.
- **Checks:** sanity/limiting-case check
- **Verdict:** PASS; confidence: medium; impact: minor
- **Assumptions/inputs:** Inputs are filtered to be positive (Sec. 2.2), λ constraints as given
- **Notes:** With $x_o > 0$ and feasibility via $\lambda_o = 1$, the optimum satisfies $\theta^* \leq 1$. A strict lower bound of 0 is not typically attained under positive inputs, so '0 to 1' is slightly imprecise but not a functional contradiction.

7. ⚠ **Interpretation of input-oriented inefficiency** (Sec. 2.3 final sentences explaining scores < 1 (p.3))
    - **Claim:** Scores less than 1 indicate inefficiency, suggesting the clinic could reduce inputs or increase outputs to improve performance.
    - **Checks:** interpretation vs formulation consistency
    - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
    - **Assumptions/inputs:** Model is input-oriented as stated
    - **Notes:** Given the presented LP, θ directly measures proportional input reduction holding outputs at least fixed; output expansion is not what θ itself quantifies. The statement is directionally true in a broad Pareto sense, but it is not the specific meaning of θ in the chosen orientation.
8. ⚠ **Zero-output ($y_o = 0$) special case behavior** (Sec. 3.5 sensitivity analysis narrative (p.7-9))
    - **Claim:** DMUs with zero live births have very low efficiency scores, and zero-output cycles substantially impact efficiency.
    - **Checks:** analytic special-case check
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** DEA model is exactly the LP given (input-oriented BCC), Zero outputs are allowed in the cleaned data
    - **Notes:** Analytically, when $y_o = 0$ the output constraint becomes nonbinding ($\sum \lambda_j y_j \geq 0$), so θ is determined solely by achievable convex-combination inputs relative to x_o; a zero-output DMU is not mathematically forced to have very low θ (it could be efficient if it is input-minimal). The paper's claim may be empirically true for their dataset, but it is not a direct mathematical implication of the LP.

Limitations

- The audit is restricted to the mathematics explicitly shown in the provided PDF text/images; there are no detailed derivations beyond the DEA LP to verify.

- No access to the referenced CSV outputs or code; therefore, the audit cannot verify whether \text{Data_Value_num} is stored as a proportion or a percent in practice—only that the paper's written definition is internally inconsistent unless clarified.
- Figures/tables are treated descriptively; numeric values and empirical claims are not checked by request.

Numerical results audit

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

Seven numerical checks were executed on dataset counts, Table 1 totals, percentage recomputations, weighted-average consistency, min/max consistency, and narrative-to-table rounding consistency. All checks passed within stated tolerances, with no execution errors.

Checked items

1. ✔ **C1** (Page 4, Section 3.1 (Dataset filtering and cleaning counts))
    - **Claim:** After initial loading of $\mathbf{1,126,080}$ records, filtering ... yielded $\mathbf{48,960}$ records. Subsequent cleaning ... and removal of $\mathbf{36}$ records ... resulted in a final dataset of $\mathbf{31,164}$ DMUs.
    - **Checks:** count_arithmetic (subtraction / implied removals)
    - **Verdict:** PASS
    - **Notes:** Implied removed_for_dea_cleaning computed as $\mathbf{17,760}$ and reconciliation $\mathbf{31,164 + 36 + 17,760 = 48,960}$ holds exactly.
2. ✔ **C2** (Page 7, Table 1 (Num DMUs by stratum) and Page 4, Section 3.1 (final dataset size))
    - **Claim:** Table 1 lists Num DMUs for $\mathbf{12}$ strata; these should sum to the stated final dataset size of $\mathbf{31,164}$ DMUs.
    - **Checks:** parts_to_total (sum of strata counts)
    - **Verdict:** PASS
    - **Notes:** Sum of the $\mathbf{12}$ Table 1 Num DMUs equals $\mathbf{31,164}$.
3. ✔ **C3** (Page 7, Table 1 (% Efficient vs Num Efficient / Num DMUs))
    - **Claim:** For each stratum in Table 1, % Efficient should equal $\mathbf{100 \times \text{Num Efficient}/\text{Num DMUs}}$ (up to rounding).
    - **Checks:** percent_recompute
    - **Verdict:** PASS
    - **Notes:** All $\mathbf{12}$ rows match $\mathbf{\text{round}(100 \times \text{Num Efficient}/\text{Num DMUs}, 2)}$ exactly; max absolute difference after rounding to $\mathbf{2}$ decimals is $\mathbf{0.00}$ percentage points.

4. ✔ **C4** (Page 6, Section 3.3 (Aggregate efficiency statistics) and Page 7, Table 1 (stratum means and counts))
    - **Claim:** Aggregate mean efficiency reported as **0.2488** should equal the DMU-count-weighted mean of the **12** stratum mean efficiencies in Table 1.
    - **Checks:** weighted_average_recompute
    - **Verdict:** PASS
    - **Notes:** Computed Table 1 DMU-weighted mean = **0.2488404698** versus reported **0.2488**; difference consistent with rounding of stratum means.
5. ✔ **C5** (Page 7, Section 3.3.1 (claimed range for mean efficiencies) vs Table 1)
    - **Claim:** Mean efficiency scores were said to range from approximately **0.17** ($>$ **40** in **2022**) to **0.34** ($<$ **35** in **2020**).
    - **Checks:** min_max_from_table
    - **Verdict:** PASS
    - **Notes:** Table 1 min mean efficiency is **0.1719** and max is **0.3434**, consistent with the narrative 'approximately **0.17**' to '**0.34**'.
6. ✔ **C6** (Page 7, Section 3.4.1 (Temporal mean efficiency trends) vs Table 1)
    - **Claim:** For $<$ **35**: **0.343** (**2020**) $\rightarrow$ **0.325** (**2021**) $\rightarrow$ **0.319** (**2022**); **35-37**: **0.258**, **0.250**, **0.261**; **38-40**: **0.214**, **0.217**, **0.235**; $>$ **40**: **0.186**, **0.178**, **0.172**.
    - **Checks:** rounded_value_consistency
    - **Verdict:** PASS
    - **Notes:** Narrative trend values equal Table 1 mean efficiencies rounded to **3** decimals for all groups/years.
7. ✔ **C7** (Page 7, Section 3.4.1 (Temporal median efficiency trends) vs Table 1)
    - **Claim:** For $<$ **35**: median efficiency was **0.250** (**2020**), **0.200** (**2021**), **0.250** (**2022**).
    - **Checks:** exact_match_to_table
    - **Verdict:** PASS
    - **Notes:** Narrative medians match the Table 1 medians for $<$ **35** in **2020**–**2022** exactly as stated.

Limitations

- Audit is restricted to numerics explicitly present in the provided PDF text; referenced CSV files and underlying NASS data are not available here.
- No numeric extraction from figures/plots was performed (per instruction), so plot-based quantitative claims cannot be checked.
- Some consistency checks (e.g., exact aggregate mean from stratum means) may show small differences due to rounding of table values; tolerances are suggested accordingly.

- Several checks require access to underlying datasets/outputs (row-level live-birth calculation and filtering, aggregate descriptive statistics including NaN checks, plot-backed distributional claims, and sensitivity-analysis outputs) and therefore remain unverified.