Draft version August 6, 2025
Typeset using LaTeX **twocolumn** style in AASTeX631

Efficiency and Performance Frontier Analysis of US ART Clinics (2020-2022) using Data Envelopment Analysis

Denario[1]

[1] *Anthropic, Gemini & OpenAI servers. Planet Earth.*

ABSTRACT

Evaluating the technical efficiency of Assisted Reproductive Technology (ART) clinics is essential for identifying best practices and ultimately improving patient outcomes. This study applies Data Envelopment Analysis (DEA) to rigorously assess the relative efficiency of U.S. ART clinics in transforming intended retrieval cycles into live births. Utilizing comprehensive clinic-level data from the 2020-2022 National ART Surveillance System (NASS) dataset, focusing on own-egg cycles, the analysis stratified clinics into distinct patient age groups (<35, 35-37, 38-40, >40). An input-oriented Banker, Charnes, Cooper (BCC-I) DEA model, assuming variable returns to scale (VRS), was then employed independently for each year and age group, defining intended retrieval cycles as the sole input and calculated live births as the sole output. Analyzing 31,164 clinic-year-age group units, the study revealed consistently low mean and median efficiency scores across all strata, with only a small fraction (2-5

Keywords: Algorithms, Computational methods, Distributed computing, Astrostatistics, Time series analysis

1. INTRODUCTION

Infertility represents a significant global health challenge, affecting an estimated one in six individuals during their lifetime. Assisted Reproductive Technology (ART) has emerged as a transformative medical intervention, offering critical pathways to parenthood. For patients undergoing ART, the ultimate and most impactful outcome is the achievement of a live birth. Given the substantial emotional, physical, and financial investments involved in ART treatments, ensuring the highest possible efficiency and performance of ART clinics is paramount. This necessitates a robust framework for evaluating how effectively clinics convert their operational efforts and resources into desired clinical outcomes.

However, objectively assessing and comparing the true performance of ART clinics presents considerable complexities. Traditional metrics, often relying solely on reported live birth rates, frequently fall short because they do not adequately account for the resources consumed or the inherent variability in patient populations. Clinics exhibit wide heterogeneity in their scale, infrastructure, patient demographics, and clinical protocols. This variability means that a high live birth rate from one clinic may be achieved with disproportionately higher inputs or with a less challenging patient cohort compared to another. Consequently, defining "efficiency" in the intricate context of ART involves understanding how well a clinic utilizes its inputs, such as the effort and resources invested in treatment cycles, to achieve optimal outputs, specifically live births, rather than merely attaining high success rates irrespective of cost or resource intensity. This intricate interplay between inputs and outputs, coupled with the heterogeneous nature of patient cohorts and operational practices, renders direct comparisons based on raw success rates misleading and highlights the limitations of conventional performance indicators.

To rigorously address these complexities, this study employs Data Envelopment Analysis (DEA), a non-parametric mathematical programming technique well-suited for evaluating the relative efficiency of a set of comparable entities, known as Decision Making Units (DMUs). Unlike traditional statistical methods that require a pre-specified functional relationship between inputs and outputs, DEA constructs an empirical "efficiency frontier" based on the observed performance of the best-performing DMUs. This frontier represents the maximum output achievable for a given set of inputs, or conversely, the minimum inputs required for a given set of outputs. By comparing each clinic to this empirically derived frontier, DEA generates a relative efficiency score, quantitatively identifying clinics operating at the frontier (technically efficient) and those operating below it (inefficient), along with their potential for improvement. For this analysis, we specifically utilize an input-oriented Banker, Charnes, Cooper (BCC-I) model, which assumes Variable Returns to Scale (VRS). This choice is particularly appropriate for healthcare settings like ART clinics, as it allows for the comparison of units operating at different scales and focuses on their ability to minimize the utilization of inputs while maintaining or increasing output levels.

Leveraging comprehensive clinic-level data from the National ART Surveillance System (NASS) dataset for the years 2020-2022, our study focuses exclusively on own-egg cycles to ensure a consistent and comparable patient population. Within this framework, we define the "Number of Intended Retrieval Cycles" as the sole input, representing the initial resource commitment and effort expended by a clinic, and the "Number of Live Births" as the sole output, representing the desired clinical outcome. A critical aspect of our methodology involves stratifying clinics into distinct patient age groups—specifically, <35, 35-37, 38-40, and >40 years. This stratification is crucial for creating more homogeneous comparison groups, as patient age is a well-established primary determinant of ART success rates and significantly influences a clinic's operational context and expected outcomes. By conducting independent DEA analyses for each year and each defined age group, we aim to capture the nuanced performance landscape within specific patient cohorts and allow for accurate peer-to-peer comparisons that reflect the inherent biological and clinical realities.

Through this rigorous application of DEA, we aim to systematically evaluate the technical efficiency of U.S. ART clinics in transforming intended retrieval cycles into live births. The analysis will reveal the distribution of efficiency scores across various strata, pinpointing the proportion of clinics operating on the efficiency frontier and quantifying the extent of performance heterogeneity within the ART landscape. We will benchmark clinics against their most efficient peers, thereby identifying best practices and potential areas for operational optimization that could be disseminated to improve overall ART outcomes. Furthermore, the study investigates temporal changes in efficiency between 2020 and 2022, offering insights into industry-wide trends and the potential impacts of evolving healthcare practices or external factors. Finally, a sensitivity analysis will explore the implications of specific data characteristics, such as the reporting of zero live births, on measured efficiency scores. Collectively, our findings will provide a critical evidence base for policymakers, clinic administrators, and patients, highlighting specific areas for targeted interventions to enhance the technical efficiency and quality of ART services, ultimately contributing to improved patient care and outcomes across the United States.

2. METHODS

2.1. *Data collection and preparation*

2.1.1. *Data source and scope*

The foundational dataset for this study was derived from the National ART Surveillance System (NASS), managed by the Centers for Disease Control and Prevention (CDC). NASS collects comprehensive data on all assisted reproductive technology (ART) cycles performed in clinics across the United States. For this analysis, we utilized clinic-level aggregate data spanning the years 2020, 2021, and 2022. To ensure comparability and focus on the core clinical process of ART, the scope of the data was specifically restricted to cycles involving patients using their own eggs, as indicated by the 'Type' column in the dataset. This exclusion of donor egg cycles mitigates the confounding effect of varying oocyte quality on clinic performance metrics.

2.1.2. *Definition of decision making units (DMUs)*

In Data Envelopment Analysis (DEA), the entities being evaluated are termed Decision Making Units (DMUs). Given the inherent variability in patient demographics and the year-to-year operational changes within clinics, each unique combination of 'ClinicId', 'Year', and 'AgeGroup' was defined as an individual DMU. This granular approach allows for more precise peer-to-peer comparisons, acknowledging that a clinic's efficiency may vary across different patient populations and over time. Clinics with insufficient data for a specific year or age group were excluded from the analysis for that particular stratum.

2.1.3. *Input and output variables*

A critical step in DEA is the precise definition of inputs and outputs. Consistent with the study's objective of evaluating how effectively clinics convert their operational efforts into desired clinical outcomes, a single input and a single output were identified for the DEA model:

- **Input: Number of Intended Retrieval Cycles.** This variable, directly extracted from the 'Cycle_Count' column for records pertaining to "% Live Births per Intended Retrieval" (identified via the 'Question' column), represents the primary resource commitment and operational effort expended by an ART clinic for a given patient cohort. It signifies the initial investment in a treatment cycle up to the point of oocyte retrieval.

- **Output: Number of Live Births.** This variable, representing the ultimate desired clinical outcome, was calculated by applying the reported percentage of live births per intended retrieval (‘Data_Value_num‘ column, typically a percentage value) to the ‘Cycle_Count‘ (Intended Retrieval Cycles). Specifically, $LiveBirths = (Data\_Value\_num/100) * Cycle\_Count$. The resulting value was then rounded to the nearest integer, as live births are discrete counts.

This input-output pairing directly quantifies the clinic's effectiveness in transforming initial treatment efforts into successful deliveries, aligning with the paper's central theme of efficiency in ART.

2.1.4. *Data preprocessing and stratification*

The raw NASS dataset underwent several preprocessing steps to prepare it for DEA:

1. **Year Filtering:** Data was filtered to include only records from 2020, 2021, and 2022.

2. **Own-Egg Cycles Filter:** Only records where the ‘Type‘ column indicated "Patients using their own eggs" were retained. This ensured a homogeneous comparison group, focusing on the intrinsic efficiency of clinics rather than variations introduced by donor gametes.

3. **Metric Identification:** The dataset was further refined to select rows where the ‘Question‘ column explicitly referred to "% Live Births per Intended Retrieval" and the ‘Breakout_Category‘ was "Age group of patient". This ensured that ‘Data_Value_num‘ represented the relevant live birth rate percentage and ‘Cycle_Count‘ represented the intended retrieval cycles for that specific metric.

4. **Age Group Stratification:** Recognizing that patient age is a dominant factor influencing ART success rates, clinics were stratified into four distinct age groups based on the ‘Breakout‘ column: <35 years, 35-37 years, 38-40 years, and >40 years. This stratification created more homogeneous comparison groups, ensuring that DEA comparisons were made between clinics treating similar patient populations, thereby allowing for a more equitable assessment of technical efficiency.

5. **Handling Missing and Invalid Data:** For DEA, all inputs and outputs must be positive values. Rows with missing, zero, or negative ‘Input_IntendedRetrievals‘ or ‘Output_LiveBirths‘ were excluded. Specifically, clinics with zero ‘Intended Retrieval Cycles‘ were removed, as they cannot serve as meaningful DMUs. While zero ‘Live Births‘ were permitted as an output, negative values were excluded. This rigorous cleaning process resulted in 31,164 clinic-year-age group units for the final analysis, ensuring the robustness and validity of the DEA model.

2.2. *Exploratory data analysis*

Prior to conducting the DEA, an exploratory data analysis (EDA) was performed on the prepared dataset. The primary objective of the EDA was to understand the distribution and characteristics of the input (‘Number of Intended Retrieval Cycles‘) and output (‘Number of Live Births‘) variables across the different strata (Year and Age Group). For each unique combination of year and age group, the following descriptive statistics were computed: the number of unique clinics (DMUs), mean, median, minimum, maximum, and standard deviation for both the input and output variables. Additionally, the mean and median of the original ‘Data_Value_num‘ (percentage live birth per intended retrieval) were analyzed to provide context on the raw success rates within each stratum. These statistics were compiled into summary tables to highlight the heterogeneity in clinic scale and performance across different patient age cohorts and over the study period. An example of the table structure generated is provided below:

Age Group	Num Clinics	Mean IR	Median IR	Min I
<35	N1	...	...	...
35-37	N2	...	...	...
38-40	N3	...	...	...
>40	N4	...	...	...

Note: IR = Intended Retrievals, LB = Live Births, LB Rate = % Live Birth per Intended Retrieval. This table structure was generated for each year of analysis.

2.3. *Data envelopment analysis (DEA) model*

2.3.1. *Model selection*

To rigorously assess the technical efficiency of ART clinics, this study employed an input-oriented Banker, Charnes, Cooper (BCC-I) DEA model. This choice is particularly apt for healthcare settings, including

ART clinics, as it assumes Variable Returns to Scale (VRS). The VRS assumption allows for the comparison of DMUs operating at different scales, which is common among ART clinics varying significantly in size and patient volume. An input-oriented model was selected because it focuses on the ability of a clinic to minimize its inputs (intended retrieval cycles) while maintaining its current level of output (live births). This perspective is highly relevant for optimizing resource utilization in a high-cost medical field like ART.

2.3.2. *Mathematical formulation*

For each DMU o (clinic-year-age group unit) being evaluated, the input-oriented BCC model solves the following linear programming problem:

Minimize θ_o

Subject to:

$$\sum_{j=1}^{N} \lambda_j x_{ij} \leq \theta_o x_{io} \qquad \forall i = 1, \ldots, M$$

$$\sum_{j=1}^{N} \lambda_j y_{rj} \geq y_{ro} \qquad \forall r = 1, \ldots, S$$

$$\sum_{j=1}^{N} \lambda_j = 1$$

$$\lambda_j \geq 0 \qquad \forall j = 1, \ldots, N$$

where:

- N is the total number of DMUs within the specific stratum (year and age group).
- x_{ij} represents the i-th input for DMU j. In this study, $M = 1$, and x_{1j} is the Number of Intended Retrieval Cycles for DMU j.
- y_{rj} represents the r-th output for DMU j. In this study, $S = 1$, and y_{1j} is the Number of Live Births for DMU j.
- x_{io} and y_{ro} are the inputs and outputs for the specific DMU o currently being evaluated.
- λ_j are the weights assigned to each DMU j in forming the virtual composite DMU.
- θ_o is the efficiency score for DMU o. A value of $\theta_o = 1$ indicates that DMU o is technically efficient and lies on the efficiency frontier, meaning it produces the maximum possible output for its given inputs or uses the minimum possible inputs for its given outputs. A value of $\theta_o < 1$ indicates inefficiency, suggesting that the DMU could potentially reduce its inputs by a factor of $(1 - \theta_o)$ while maintaining its current output level.
- The constraint $\sum_{j=1}^{N} \lambda_j = 1$ is the convexity constraint specific to the BCC model, which ensures variable returns to scale.

2.3.3. *Computational implementation*

The DEA model was implemented using Python, leveraging the 'scipy.optimize.linprog' function, which solves linear programming problems. For each unique stratum (combination of 'Year' and 'AgeGroup'), the DEA was executed independently. This involved:

1. Identifying all DMUs belonging to the current stratum.
2. Extracting their respective 'Input_IntendedRetrievals' and 'Output_LiveBirths' values.
3. For each DMU o within the stratum, the linear programming problem described above was solved to compute its efficiency score θ_o.

To manage the computational load associated with analyzing thousands of DMUs across multiple strata, the DEA calculations were parallelized using Python's 'multiprocessing' library. This allowed for concurrent computation of efficiency scores across different DMUs, significantly reducing the overall processing time. The resulting efficiency scores for each clinic-year-age group unit were then stored for subsequent analysis.

2.4. *Post-DEA analysis*

2.4.1. *Efficiency score analysis*

Following the computation of efficiency scores for all DMUs, a comprehensive analysis of their distribution was performed. For each year and age group stratum, descriptive statistics of the 'Efficiency_Score' were calculated, including mean, median, minimum, maximum, and standard deviation. Furthermore, the number and % of technically efficient clinics (those with an efficiency score of approximately 1, specifically $\geq$ 0.9999 to account for floating-point precision) were identified. This analysis provided a quantitative overview of the performance landscape within each distinct patient cohort.

2.4.2. *Temporal and age-group analysis*

To understand the dynamics of efficiency over time and across different patient demographics, the aggregated efficiency statistics were further analyzed. We examined the trends in mean and median efficiency scores across the years 2020, 2021, and 2022 for each age group. This temporal analysis aimed to identify any improvements, deteriorations, or stability in overall ART clinic efficiency. Concurrently, a cross-sectional analysis compared the mean and median efficiency scores among the different age groups within each year, highlighting the impact of patient age on measured technical efficiency.

2.4.3. *Sensitivity analysis for zero live births*

A sensitivity analysis was conducted to assess the impact of reporting zero live births on a clinic's measured efficiency score. Clinics that reported zero live births for a specific year and age group were included in the primary DEA model. The efficiency scores of these clinics were then specifically examined to determine how their null output influenced their position relative to the efficiency frontier. This direct observation allowed us to quantify the extent to which reporting no live births, despite performing intended retrieval cycles, affected a clinic's calculated technical efficiency, providing insights into the robustness of the efficiency measurement for such cases.

3. RESULTS

The results section details the findings from the Data Envelopment Analysis (DEA) applied to U.S. Assisted Reproductive Technology (ART) clinics for the years 2020-2022. The analysis focused on evaluating the technical efficiency of clinics in converting intended egg retrieval cycles into live births, meticulously stratified by patient age group to ensure comparable peer groups.

3.1. *Dataset and DEA Input-Output Specification*

The analysis was conducted using comprehensive clinic-level data from the Centers for Disease Control and Prevention's (CDC) National ART Surveillance System (NASS) Final ART Success Rates dataset, covering reporting years 2020, 2021, and 2022. To maintain focus on the core clinical process and ensure homogeneity in comparisons, the dataset was rigorously filtered to include only cycles where patients utilized their own eggs. The specific metric extracted from NASS for this study was "What percentage of intended egg retrievals resulted in live-birth deliveries?", further stratified by patient age groups: < 35, $35 - 37$, $38 - 40$, and > 40 years.

In line with the methodology described, each Decision Making Unit (DMU) for the DEA model was defined as a unique combination of a clinic, a reporting year, and a patient age group. An input-oriented Banker, Charnes, Cooper (BCC-I) DEA model, assuming variable returns to scale (VRS), was employed. This model's suitability for healthcare settings, as highlighted in the Introduction, allows for the evaluation of clinics operating at various scales while focusing on minimizing inputs to achieve a given output. The single input and output for the DEA model were specified as follows:

- **Input (X): Number of Intended Retrieval Cycles.** This variable represents the initial resource commitment and operational effort expended by an ART clinic for a specific patient cohort within a given year. It directly corresponds to the 'Cycle_Count' for "Live Births per Intended Retrieval" in the NASS dataset.
- **Output (Y): Number of Live Births.** This variable, the ultimate desired clinical outcome, was derived by multiplying the reported percentage of live births per intended retrieval ('Data_Value_num') by the 'Input_IntendedRetrievals' and rounding the result to the nearest whole number. This calculation directly quantifies the success rate of a clinic's efforts in achieving live births.

After initial data loading and comprehensive filtering for the specified years, own-egg cycles, relevant questions, and age breakout categories, a robust dataset of 31,164 DMUs was obtained for the final analysis. This rigorous preprocessing, including the exclusion of national summary data and records with invalid inputs/outputs, ensured the integrity and validity of the DEA computations.

3.2. *Descriptive Analysis of DEA Inputs and Outputs*

An exploratory data analysis (EDA) was performed to characterize the input and output variables across the various strata (Year and Age Group). This initial assessment provided crucial context for understanding the operational landscape of U.S. ART clinics. The number of DMUs per stratum was substantial, ranging from approximately 2,400 to 2,800 units across the years and age groups, ensuring the robustness of the DEA frontier estimation.

3.2.1. *Input: Intended retrievals*

The 'Input_IntendedRetrievals' variable exhibited considerable variability across DMUs. Mean values for intended retrievals showed a slight upward trend from 2020 to 2022 for most age groups, suggesting a general increase in activity or reporting. For instance, in the < 35 age group, the mean intended retrievals increased from 113.2 in 2020 to 133.8 in 2022. However, median values were consistently much lower than the means (e.g., 43.0 vs. 113.2 for < 35 in 2020). This substantial difference between mean and median, coupled with large standard deviations and exceptionally high maximum values (e.g., 5021 cycles for a single DMU), indicates a highly right-skewed distribution of intended retrievals. This skewness, reflecting wide heterogeneity in clinic scale and operational intensity, is clearly illustrated in the histograms and boxplots presented in Figure 1 and Figure 2, respectively. These figures demonstrate that while many clinics perform a relatively small number of cycles, a few high-volume clinics significantly influence the overall average.

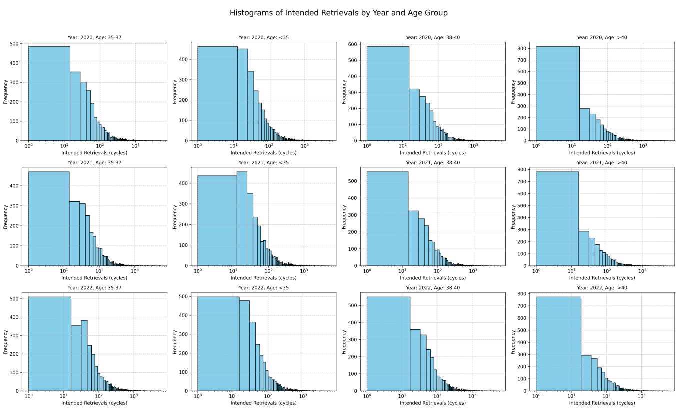

Figure 1. Histograms of Intended Retrievals (cycles) for U.S. Assisted Reproductive Technology clinics, stratified by year and patient age group. These distributions are highly right-skewed, demonstrating that most clinics perform a relatively low number of intended retrievals, while a smaller number perform substantially more. This variability reflects the input data characteristics for the Data Envelopment Analysis.

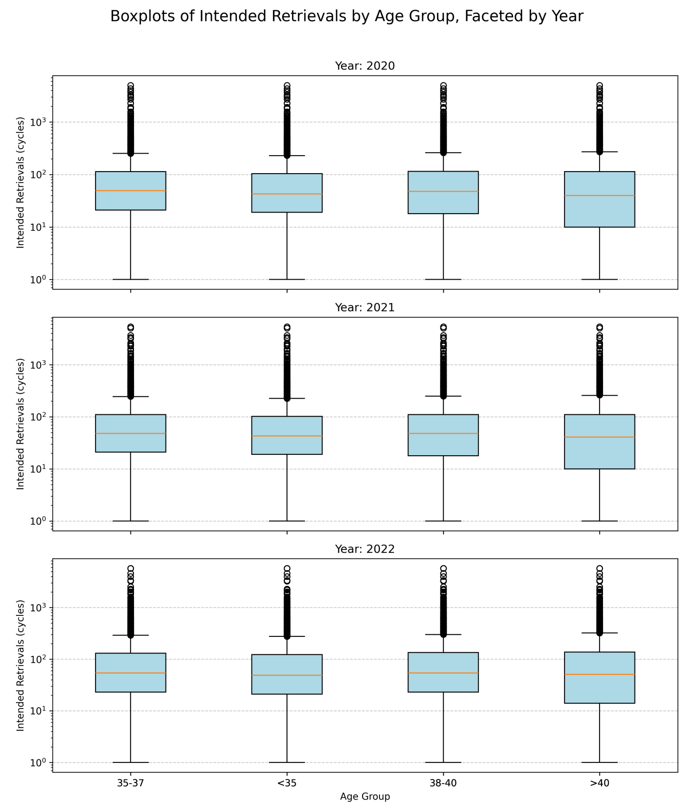

Figure 2. Boxplots display the distribution of intended egg retrieval cycles (input) by patient age group and year. The distributions are consistently right-skewed, characterized by a concentration of clinics with lower cycle counts and a long tail of high-volume outliers across all strata.

3.2.2. *Output: Live births*

The ‘Output_LiveBirths‘ variable similarly displayed wide variation and a strong right-skewness. Mean live births generally mirrored the trends in inputs, with slight increases observed for younger age groups over the study period. A critical observation was the frequent occurrence of a median ‘Output_LiveBirths‘ of 0.0, particularly for older age groups (e.g., the > 40 age group reported a median of 0.0 across all three years). This finding is highly significant, indicating that for at least half of the DMUs within these strata, no live births were reported despite the execution of intended retrieval cycles. This prevalence of zero outputs is a key factor influencing the efficiency scores, as a non-zero input with a zero output inherently indicates a high degree of inefficiency in DEA. The distributions of live births, highlighting this skewness and the prevalence of zero outcomes, are presented in the histograms in Figure 3 and the boxplots in Figure 4.

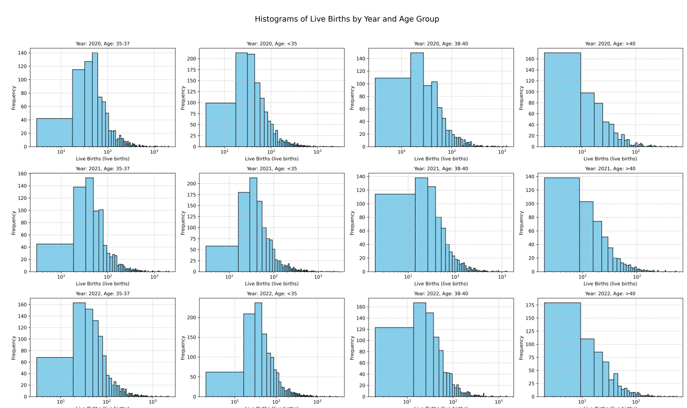

Figure 3. This figure presents histograms showing the frequency distribution of live births per decision-making unit (clinic-year-age group combination) from 2020 to 2022, stratified by patient age. The distributions exhibit strong right-skewness, with a high frequency of units reporting zero or very few live births, particularly notable in older age groups. This characteristic of the output data significantly influences the calculated technical efficiency.

3.2.3. *Live birth rate per intended retrieval*

The original reported ‘Data_Value_num‘ (percentage live birth per intended retrieval), used to calculate

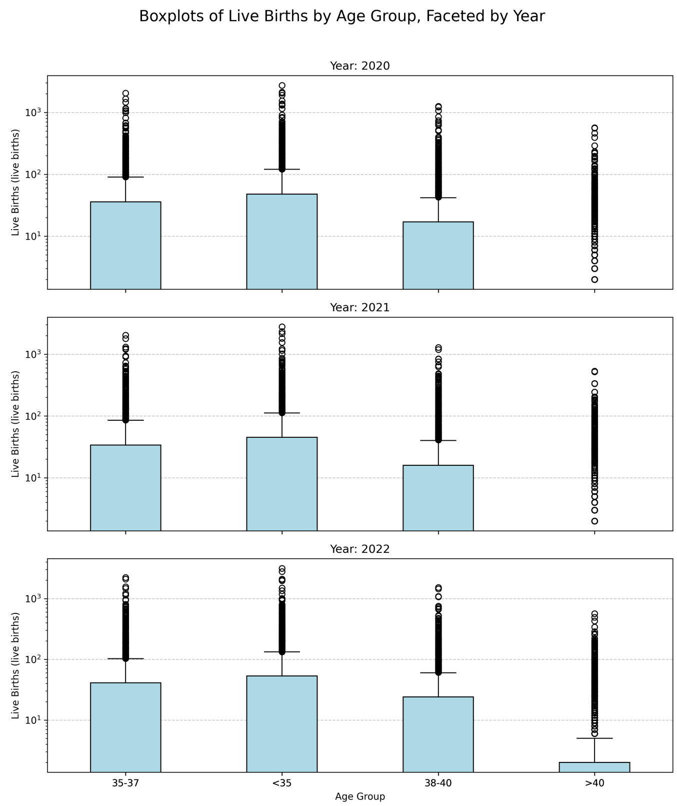

Figure 4. Boxplots illustrating the distribution of live births by patient age group for years 2020-2022, presented on a logarithmic scale. The plots reveal a pronounced right-skewness in the number of live births, with a high frequency of zero outcomes, particularly for older patient age groups. This characteristic of the output data significantly influences the calculated DEA efficiency scores.

'Output_LiveBirths', confirmed well-established biological trends. The mean live birth rate was highest for the < 35 age group (e.g., 22.46% in 2020) and systematically decreased with increasing patient age, dropping to as low as 1.95% for the > 40 age group in 2020. Consistent with the 'Output_LiveBirths' data, median live birth rates were frequently 0.0%, especially for the '35-37', '38-40', and '>40' age groups across all years. This reinforces the observation that a substantial proportion of DMUs, particularly those treating older patients, reported no successful live births for the specific metric analyzed. The distributions of these rates are shown in Figure 5 and Figure 6, clearly indicating the prevalence of low or zero live birth rates, especially for older patients. This characteristic of the input data for DEA is anticipated to significantly influence the efficiency scores, as discussed in the subsequent sections.

3.3. *Clinic Efficiency Scores*

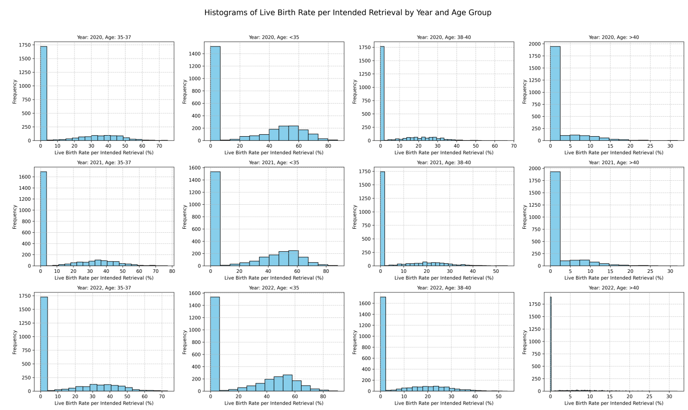

Figure 5. Histograms illustrate the distribution of Live Birth Rate per Intended Retrieval for U.S. ART clinics from 2020 to 2022, stratified by patient age group. The distributions are consistently right-skewed, with a substantial number of clinics reporting zero or very low live birth rates, particularly evident in older patient age categories (35-37, 38-40, >40). Conversely, the '<35' age group displays a more dispersed distribution, indicating higher live birth rates. This prevalence of low or zero live birth rates, especially for older patients, is a key characteristic of the input data that significantly influences the calculated Data Envelopment Analysis efficiency scores.

The BCC-I DEA model successfully computed efficiency scores for all 31,164 DMUs. An efficiency score of 1.0 indicates that a DMU operates on the efficiency frontier, representing best observed practice for its given inputs and outputs within its peer group. Scores below 1.0 denote relative inefficiency.

The aggregate statistics for all computed efficiency scores across all years and age groups revealed a mean efficiency of 0.2488, a median of 0.1000, and a standard deviation of 0.2804, with scores ranging from 0.0004 to 1.0000. The strikingly low mean and median efficiency scores suggest that, on average, U.S. ART clinics operated considerably below the theoretical efficiency frontier established by their most efficient peers during the study period. This indicates substantial room for improvement in resource utilization relative to outcomes.

3.3.1. *Stratum-specific efficiency score analysis*

A detailed breakdown of efficiency scores by year and age group is presented in Table 1.

As detailed in Table 1, mean efficiency scores remained consistently low across all strata, generally ranging from approximately 0.17 for the > 40 age group in 2022 to 0.34 for the < 35 age group in 2020. Median efficiency scores were even lower than the means, particularly for older age groups (e.g., 0.0714 for > 40 in 2022), which is indicative of a pronounced right-skewness in the distribution of efficiency scores within each stratum. This

Table 1. Summary of DEA Efficiency Scores by Year and Patient Age Group

Year	Age Group	Num DMUs	Mean Efficiency	Median Efficiency	Min Efficiency	Max Efficiency	Std Efficiency
2020	< 35	2752	0.3434	0.2500	0.0007	1.0	0.3161
2020	35-37	2548	0.2577	0.0769	0.0013	1.0	0.2921
2020	38-40	2524	0.2139	0.0807	0.0013	1.0	0.2585
2020	> 40	2464	0.1859	0.0833	0.0018	1.0	0.2458
2021	< 35	2720	0.3247	0.2000	0.0004	1.0	0.3040
2021	35-37	2496	0.2504	0.0769	0.0004	1.0	0.2851
2021	38-40	2491	0.2171	0.0833	0.0004	1.0	0.2537
2021	> 40	2441	0.1783	0.0769	0.0004	1.0	0.2386
2022	< 35	2849	0.3188	0.2500	0.0025	1.0	0.2934
2022	35-37	2699	0.2612	0.1000	0.0004	1.0	0.2835
2022	38-40	2640	0.2353	0.1000	0.0004	1.0	0.2671
2022	> 40	2540	0.1719	0.0714	0.0004	1.0	0.2339

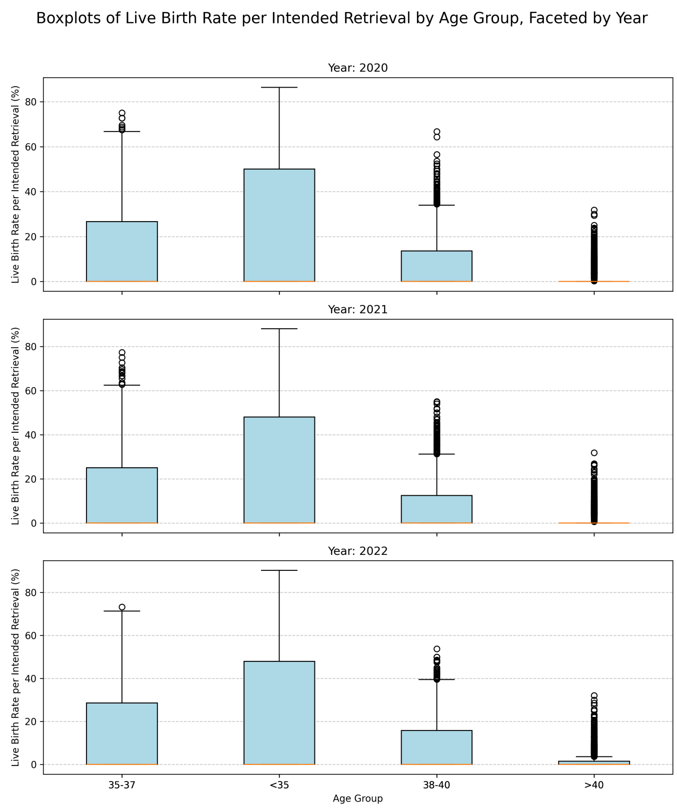

Figure 6. Boxplots displaying the distribution of live birth rates per intended retrieval for U.S. Assisted Reproductive Technology clinics from 2020 to 2022, stratified by patient age group. The live birth rates consistently decrease with increasing patient age. A notable characteristic of the input data for efficiency analysis is the high prevalence of zero live births, particularly for older age groups.

distribution pattern, showing a large number of DMUs operating at very low efficiency levels, while a smaller subset achieves higher scores, is clearly visualized in the histograms in Figure 7. This wide dispersion in efficiency scores points to significant heterogeneity in performance among U.S. ART clinics.

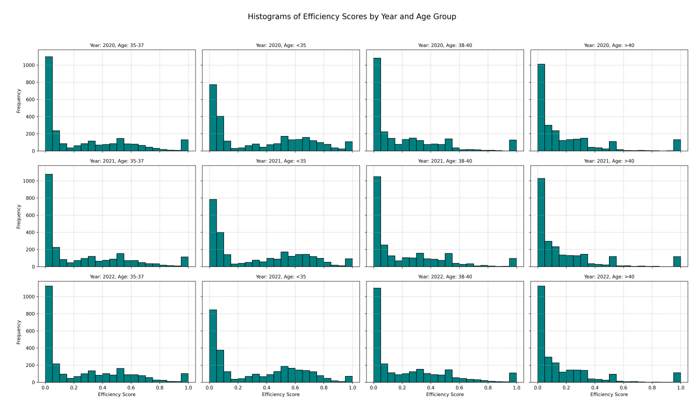

Figure 7. Histograms display the distribution of Data Envelopment Analysis (DEA) technical efficiency scores for U.S. Assisted Reproductive Technology (ART) clinics, stratified by year and patient age group (2020-2022). These distributions are consistently right-skewed, showing a high frequency of low efficiency scores and a small peak at the efficiency frontier (score of 1.0). This indicates that most clinics operate below best practice in converting intended egg retrievals into live births. Efficiency generally decreases with increasing patient age, with distributions for older age groups concentrated more heavily at lower values.

The proportion of DMUs operating on the efficiency frontier (those with a score of 1.0, or specifically ≥ 0.9999 to account for floating-point precision) was remarkably small across all strata. Typically, only 2% to 5% of clinics achieved technical efficiency. For example, in 2022, this percentage varied from 2.28% for the < 35 age group to 4.29% for the > 40 age group. This small fraction of fully efficient clinics indicates that the vast majority of ART clinics in the U.S. have substantial potential to improve their efficiency in converting intended retrieval cycles into live births by adopting best practices observed among their peers.

3.4. *Temporal and Age-Group Trends in Efficiency*

The analysis further explored how efficiency scores varied over time and across different patient demographics.

3.4.1. *Temporal trends (2020-2022)*

Temporal trends in mean efficiency scores for each age group from 2020 to 2022 are illustrated in Figure 8. For the < 35 age group, mean efficiency showed a slight decline from 0.343 in 2020 to 0.319 in 2022. In contrast, the $35 - 37$ age group exhibited relative stability in mean efficiency, hovering around 0.25-0.26 throughout the period. The $38 - 40$ age group experienced a modest increase in mean efficiency, from 0.214 in 2020 to 0.235 in 2022, while the > 40 age group saw a slight decrease from 0.186 to 0.172 over the same period. These year-over-year changes in mean efficiency were generally minor across all age groups, suggesting no substantial widespread shifts or significant improvements/deteriorations in overall clinic efficiency during the observed three-year period. Median efficiency scores also showed varied but generally minor changes, reinforcing the conclusion that the typical DMU's performance relative to the frontier did not change dramatically over time. This stability implies that the underlying factors influencing efficiency remained relatively consistent during this period.

3.4.2. *Age-group trends in efficiency*

A consistent and pronounced pattern emerged regarding the relationship between patient age group and clinic efficiency across all three years: DMUs associated with younger patient age groups consistently demonstrated higher mean and median efficiency scores. As patient age increased through the $35-37$, $38-40$, and > 40 categories, both mean and median efficiency scores systematically decreased. The > 40 age group consistently exhibited the lowest average efficiency scores. This trend is clearly visible in the violin plots presented in Figure 9, where the density for older age groups is more compressed towards lower efficiency values. This pattern suggests that achieving high relative efficiency is inherently more challenging, or performance is more variable, when treating older patient populations, even when compared against peers treating the same age group. This aligns with the known biological reality of declining fertility with age, indicating that clinics face greater inherent challenges in achieving live births from intended retrieval cycles in older patients. However, even within these more challenging cohorts, there remains a frontier of best performance against which other clinics can be benchmarked.

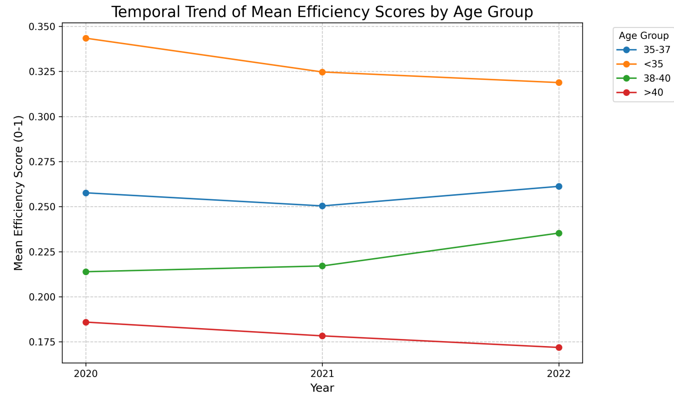

Figure 8. This figure illustrates the mean technical efficiency scores of U.S. Assisted Reproductive Technology (ART) clinics from 2020 to 2022, stratified by patient age group. It shows that mean efficiency consistently decreases with increasing patient age, with the < 35 group maintaining the highest scores and the > 40 group the lowest. Temporal changes in mean efficiency were generally minor across all age groups during this period, suggesting stable overall clinic performance.

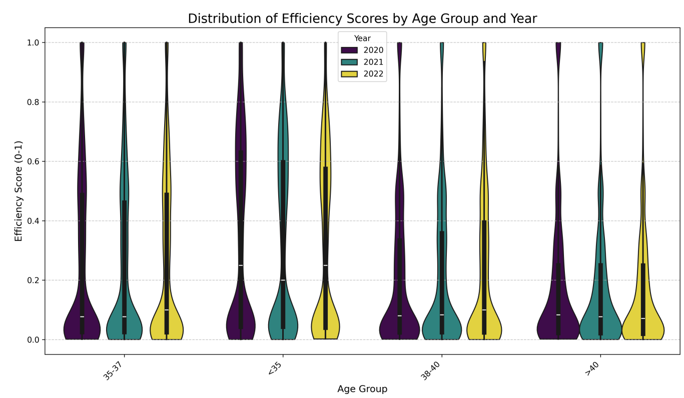

Figure 9. Violin plots display the distribution of Data Envelopment Analysis (DEA) efficiency scores for U.S. Assisted Reproductive Technology (ART) clinics from 2020 to 2022, stratified by patient age group. The distributions show that most clinics operate with low efficiency, concentrated near zero, while a small proportion achieve scores of 1.0. Efficiency generally decreases with increasing patient age, with younger age groups exhibiting distributions shifted towards higher scores. Minor temporal changes are observed within each age group.

3.5. *Impact of zero live births on efficiency (sensitivity analysis)*

A sensitivity analysis was performed to specifically assess the impact of DMUs reporting zero live births (Output_LiveBirths = 0) on their calculated efficiency scores. This analysis is crucial given the high prevalence of zero outputs observed, particularly in older age groups. Figure 10 visually compares the efficiency score distributions for DMUs with zero live births versus those with at least one live birth.

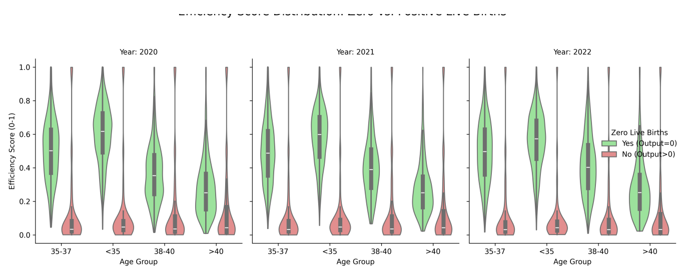

Figure 10. Violin plots showing the distribution of technical efficiency scores for U.S. ART clinics from 2020 to 2022, separated by patient age group and whether the clinic-year-age group (DMU) reported zero live births (green) or positive live births (red). DMUs with zero live births consistently exhibit efficiency scores concentrated near zero across all years and age groups. In contrast, DMUs with positive live births show distributions shifted towards higher efficiency values, indicating that achieving at least one live birth is a critical determinant of higher technical efficiency.

A pronounced and consistent difference in efficiency scores was observed between these two groups. As depicted in Figure 10, DMUs reporting zero live births consistently exhibited very low mean and median efficiency scores across all strata. For instance, in 2020 for the $35-37$ age group, DMUs with zero output had a mean efficiency of 0.139 and a median of 0.034. In stark contrast, DMUs reporting positive live births for the same stratum had substantially higher mean and median efficiency scores (mean 0.503, median 0.503). This pattern was consistent across all years and age groups. The distributions for DMUs with zero live births were heavily concentrated at the lower end of the efficiency scale (near zero), whereas those with positive live births, while still skewed, were markedly shifted towards higher efficiency values.

This finding critically underscores that a primary determinant of a clinic's low measured efficiency score in this DEA model is its failure to achieve any live births from the intended retrieval cycles for a specific DMU (clinic-year-age group unit). As an input-oriented DEA model, a DMU producing zero output from a non-zero input will inherently be assessed as highly inefficient, especially when compared to peers that manage to achieve positive outputs from similar input levels. This highlights that the ability to achieve at least one live birth is a fundamental threshold for demonstrating technical efficiency within this framework, and clinics struggling to reach this threshold will be identified as having the greatest potential for improvement.

3.6. *Benchmarking and implications*

The DEA results provide a robust framework for identifying benchmark performance and areas for potential operational optimization within the U.S. ART sector. The DMUs achieving an efficiency score of 1.0 represent the best observed practices within their respective peer groups (defined by year and patient age). The consistently low percentage of such fully efficient DMUs (typically 2%-5%) across all strata implies that a substantial majority of ART clinics operate below the observed benchmark. This indicates considerable opportunities for performance enhancement across the sector.

The wide distribution of efficiency scores and the low overall averages suggest significant heterogeneity in the performance of U.S. ART clinics in converting intended retrievals into live births, even after accounting for the critical confounding factor of patient age. The profound impact of zero live births on efficiency scores further suggests that a key area for improving overall efficiency metrics lies in addressing the underlying clinical and operational factors contributing to cycles that result in no live births. These findings collectively highlight the potential for ART clinics to improve their technical efficiency by studying and adopting the practices of their more efficient peers, thereby optimizing resource utilization and ultimately improving patient outcomes. The persistent challenge of achieving live births in older patient populations is also quantified, underscoring the need for tailored strategies and resource allocation for these cohorts.

4. CONCLUSIONS

4.1. *Overview of the study*

Infertility treatments via Assisted Reproductive Technology (ART) represent significant physical, emotional, and financial investments for patients. Evaluating the true performance of ART clinics, beyond simple success rates, is complex due to the inherent heterogeneity in clinics and patient populations and the need to account for resources consumed. This study addressed this challenge by applying Data Envelopment Analysis (DEA) to rigorously assess the technical efficiency of U.S. ART clinics. The primary objective was to quantify how effectively clinics transform their operational efforts, specifically intended retrieval cycles, into the desired outcome of live births, while accounting for the critical confounding factor of patient age.

4.2. *Data and methods*

The analysis leveraged comprehensive clinic-level data from the 2020-2022 National ART Surveillance System (NASS) dataset, focusing exclusively on own-egg cycles to ensure comparability. Each Decision Making Unit (DMU) was defined as a unique clinic-year-age group combination, leading to 31,164 DMUs. The sole input to the DEA model was the "Number of Intended Retrieval Cycles," representing resource commitment, and the sole output was the "Number of Live Births," calculated from reported success rates. An input-oriented Banker, Charnes, Cooper (BCC-I) DEA model, assuming Variable Returns to Scale (VRS), was employed independently for four distinct patient age groups (<35, 35-37, 38-40, >40) and for each year. A sensitivity analysis was also conducted to investigate the impact of reporting zero live births on efficiency scores.

4.3. *Key findings*

The DEA revealed consistently low mean and median efficiency scores across all patient age groups and years, indicating that, on average, U.S. ART clinics operate significantly below their theoretical efficiency frontier. Only a small fraction, typically between 2% and 5%, of the analyzed clinic-year-age group units were found to be technically efficient, meaning the vast majority have considerable room for improvement. A clear and consistent trend emerged where efficiency systematically decreased with increasing patient age, reflecting the biological realities of declining fertility. While some minor year-over-year fluctuations were observed, temporal changes in efficiency between 2020 and 2022 were generally minimal across all age groups, suggesting relative stability in the overall efficiency landscape during this period. Furthermore, the sensitivity analysis demonstrated that reporting zero live births for a given clinic-year-age group unit substantially reduced its measured efficiency score, highlighting the critical importance of achieving at least one live birth for demonstrating technical efficiency within this framework.

4.4. *Implications and future directions*

This study provides critical insights into the performance landscape of U.S. ART clinics. The observed low average efficiency and the small proportion of efficient clinics underscore substantial heterogeneity in performance and considerable opportunities for enhancing the technical efficiency of ART services. By identifying clinics operating on the efficiency frontier, this analysis provides a robust framework for benchmarking best practices. Inefficient clinics can learn from their more efficient peers within their specific patient age cohorts, adopting optimized protocols and resource allocation strategies to improve their input-output conversion. The consistent decline in efficiency with increasing patient age quantifies the inherent challenges in treating older populations, suggesting that tailored approaches and potentially different efficiency benchmarks may be valuable for these cohorts. The significant impact of zero live births on efficiency scores highlights a key area for improvement: clinics must focus on strategies to maximize live birth outcomes from all intended retrieval cycles. Future research could explore the specific operational, clinical, and structural factors that differentiate efficient clinics from inefficient ones, providing actionable insights for quality improvement initiatives. Furthermore, expanding the model to include additional inputs (e.g., staff time, laboratory resources) and outputs (e.g., successful embryo transfers) could offer a more comprehensive view of ART clinic efficiency. Ultimately, improving the technical efficiency of ART clinics holds the potential to optimize resource utilization, reduce costs, and, most importantly, improve patient outcomes and access to successful fertility treatments across the United States.