Draft version June 16, 2025
Typeset using LaTeX **twocolumn** style in AASTeX631

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

Denario[1]

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

ABSTRACT

Understanding the variability in Assisted Reproductive Technology (ART) clinic performance is crucial for patients and practitioners, particularly during periods of potential disruption such as the COVID-19 pandemic (2020-2022). This study aimed to characterize the year-to-year variability in key U.S. ART clinic success and efficiency metrics between 2020 and 2022 and identify associated clinic-level factors. Utilizing clinic-level data from the National ART Surveillance System (NASS) for these years, we analyzed variability in metrics including live birth rates per retrieval and average retrievals/transfers per live birth, stratified by patient age group and egg source (own vs. donor). Variability was quantified using the Coefficient of Variation and Standard Deviation for each clinic across the three-year period. Associations between this variability and clinic volume (average cycle count) and geographic location (state) were explored using Spearman correlations and Ordinary Least Squares regression models. While limitations precluded analysis of live birth per transfer and a significant anomaly was noted in 2022 donor egg reporting, analysis of available metrics revealed substantial year-to-year variability in clinic performance and efficiency. Counterintuitively, higher clinic volume was consistently associated with higher relative and absolute variability in own-egg and donor-egg success rates, while showing negative associations with variability in some efficiency metrics. Geographic location demonstrated some state-specific associations with variability, but these were not uniform across all metrics or patient groups, and overall, clinic volume and state explained only a modest portion of the observed variability. These findings highlight complex dynamics in ART clinic performance variability during the pandemic era, suggesting that higher volume clinics may experience larger fluctuations in success rates, and underscore the importance of considering clinic characteristics and data reporting challenges in national ART surveillance.

Keywords: Computational methods, Multivariate analysis, F test, Linear regression, Regression

1. INTRODUCTION

Assisted Reproductive Technology (ART) represents a critical advancement in addressing infertility, providing pathways to parenthood for millions globally. The success and efficiency of ART procedures, commonly quantified by metrics such as live birth rates per cycle and the average number of treatment cycles required for a successful outcome, are paramount concerns for both patients and healthcare providers. However, it is widely acknowledged that performance can exhibit considerable variation among different ART clinics. Comprehending the nature, magnitude, and underlying factors contributing to this inter-clinic variability is essential for empowering patients to make informed decisions, guiding quality improvement initiatives within clinics, and facilitating effective public health surveillance of ART outcomes.Characterizing ART clinic performance variability is inherently challenging due to its multifactorial nature. Success rates are influenced by a complex interplay of patient-specific attributes (e.g., age, diagnosis), clinic-specific factors (e.g., experience, protocols, laboratory practices, volume of cycles performed), and external environmental conditions. Analyzing these influences requires access to comprehensive, standardized data collected consistently across a large number of clinics over extended periods.

Furthermore, healthcare systems, including elective medical services like ART, are susceptible to disruptions from external shocks. The period between 2020 and 2022 was profoundly marked by the global COVID-19 pandemic, which imposed unprecedented challenges such as temporary clinic closures, supply chain interruptions, shifts in patient behavior, and the necessity for

rapid adaptation of clinical workflows and safety protocols (Irons & Raftery 2024; Oveson et al. 2025). These disruptions could plausibly impact clinic operations and performance metrics, potentially altering or exacerbating existing patterns of variability.

Investigating ART clinic performance specifically during this turbulent period offers a unique opportunity to assess the dynamics of variability under stress and identify clinic-level characteristics associated with greater resilience or vulnerability to such external pressures (Harris et al. 2024).

Despite the recognized importance of variability, few studies have systematically characterized the *year-to-year* fluctuations in ART clinic performance, particularly in the context of a major systemic shock like the COVID-19 pandemic (Parker et al. 2024). The specific ways in which pandemic-related disruptions might have affected different performance metrics (success vs. efficiency) and different patient groups (e.g., those using own eggs vs. donor eggs, different age groups), and whether certain clinic characteristics influenced the degree of this variability, remain underexplored (Caldera et al. 2024).

This study aims to bridge this gap by systematically characterizing the year-to-year variability in key U.S. ART clinic success and efficiency metrics during the COVID-19 pandemic years (2020, 2021, and 2022).

We attempt to address this problem by leveraging clinic-level data from the National ART Surveillance System (NASS) for the specified period (Bandara et al. 2020). We analyze fluctuations in key performance indicators, including success rates such as the percentage of cycles resulting in a live birth per retrieval, and efficiency metrics such as the average number of retrievals or transfers required per live birth. We examine this variability at the clinic level, stratifying our analysis to account for important patient subgroups, specifically differentiating between cycles utilizing patients' own eggs versus donor eggs, and further stratifying own-egg cycles by patient age group.

To quantify the year-to-year variability for each clinic within these specific patient strata across the three-year period, we calculate standard statistical measures: the Coefficient of Variation (CV) and the Standard Deviation (SD). Beyond merely describing the extent of this variability, we investigate whether readily available clinic-level factors are associated with the degree of performance fluctuation observed during this pandemic era. Specifically, we explore the relationships between the calculated variability measures and clinic volume, measured by the average cycle count over the 2020-2022 period for the relevant stratum, and geographic location, indicated by the state in which the clinic is located.

Through this systematic approach, utilizing robust data and statistical methods to quantify clinic-specific, stratum-specific year-to-year variability (Oganisian et al. 2024) and analyze its correlates, this study seeks to provide novel insights into the complex dynamics of ART clinic performance during a period of significant external disruption (Montoya et al. 2025).

By characterizing these patterns and identifying clinic-level factors associated with variability (Oganisian et al. 2024), we aim to contribute to a better understanding of how ART clinics navigated the challenges of the pandemic and inform future efforts in ART surveillance, quality assessment, and potentially clinic management strategies to mitigate performance fluctuations (**?**).

2. METHODS

2.1. *Data Source and Study Population*

This study utilized publicly available, clinic-level aggregate data from the National ART Surveillance System (NASS), maintained by the Centers for Disease Control and Prevention (CDC). NASS collects data on nearly all ART cycles performed in the United States. For the purpose of characterizing performance variability during the COVID-19 pandemic, we focused on data reported for the years 2020, 2021, and 2022. The dataset represents clinic-level aggregated outcomes stratified by various factors, including patient age group and egg source, rather than individual patient data. The initial dataset was loaded from a comma-separated values (CSV) file (‘art_data_2020_2024.csv‘) and filtered to include only records pertaining to reporting years 2020, 2021, and 2022.

2.2. *Outcome Measures and Clinic Characteristics*

The primary outcome measures were key ART clinic performance and efficiency metrics reported in NASS. These metrics were identified within the dataset by combining information from columns such as ‘Topic‘, ‘Question‘, ‘Type‘, ‘Filter‘, ‘Breakout_Category‘, and ‘Breakout‘ (Abbasian et al. 2024; Chen et al. 2025). The specific metrics analyzed were:

- Percentage of live births per intended retrieval (% Live Birth per Intended Retrieval)
- Percentage of live births per transfer (% Live Birth per Transfer)
- Percentage of live births per actual retrieval (% Live Birth per Actual Retrieval)

- Average number of transfers per intended retrieval (Avg Transfers per Intended Retrieval)
- Average number of intended retrievals per live birth (Avg Intended Retrievals per Live-Birth)

These metrics were extracted from the 'Data_Value_num' column.To account for important patient subgroups known to influence ART outcomes (Silver et al. 2020), analyses were stratified by egg source and patient age. Egg source was differentiated based on 'Topic' or 'Type' values indicating "patients using their own eggs" or "patients using donor eggs," creating an 'EggSource' variable ('Own' or 'Donor'). For cycles using own eggs, patient age groups (<35, 35-37, 38-40, >40) were identified from the 'Breakout' column where 'Breakout_Category' was 'Age group of patient', standardizing age group labels. For donor egg cycles, a general 'All Ages Donor' stratum was used.Clinic-level characteristics explored as potential correlates of variability were clinic volume and geographic location. Clinic volume for a given stratum (clinic, egg source, age group) in a given year was represented by the 'Cycle_Count' reported for that specific stratum and year. Geographic location was indicated by the state in which the clinic was located, using the 'LocationAbbr' column (Dai et al. 2025). Each unique clinic was identified by its 'ClinicId'.

2.3. *Data Processing and Structuring*

The raw NASS data was in a long format, with multiple rows per clinic-year combination, each corresponding to a specific metric or stratum. To enable clinic-level analysis of metrics over the three-year period (2020-2022), the data was restructured into a wide format. This involved pivoting the data such that each row represented a unique clinic-stratum combination ('ClinicId', 'EggSource', 'AgeGroup'), and columns contained the values for each of the selected performance and efficiency metrics ('Data_Value_num'), 'Cycle_Count', and 'LocationAbbr' for each of the three years (2020, 2021, 2022). Specifically, for each combination of 'ClinicId', 'Year', 'EggSource', 'AgeGroup', and 'LocationAbbr', relevant metric values and the corresponding 'Cycle_Count' were extracted. A 'pivot_table' operation was used with '['ClinicId', 'Year', 'EggSource', 'AgeGroup', 'LocationAbbr']' as the index and metric identifiers (derived from descriptive columns) as columns, populated by 'Data_Value_num' and 'Cycle_Count'.Data cleaning involved converting metric values to a consistent numeric format (float), ensuring percentages were represented uniformly (0-100 scale) (Goyle et al. 2023; Bendinelli et al. 2025). Missing values ('NaN') in metric columns were retained as such, as they indicate years where a clinic did not report data for a specific metric within a given stratum (Lee et al. 2021; Goyle et al. 2023). No imputation of missing metric values was performed.

2.4. *Quantification of Year-to-Year Variability*

Year-to-year variability in each performance and efficiency metric was quantified for each clinic within each defined stratum ('EggSource', 'AgeGroup') (Tran et al. 2022; Jones et al. 2025). To calculate variability, a clinic-stratum had to have reported data for the specific metric in at least two of the three years (2020, 2021, 2022) (Guan et al. 2022; Bedi et al. 2025). Clinics or strata with data for fewer than two years for a given metric were excluded from the variability calculation for that metric (Bedi et al. 2025).For each eligible clinic-stratum and for each metric, two measures of variability were calculated across the available years (2020, 2021, 2022):

- **Coefficient of Variation (CV):** Calculated as $\left(\frac{\text{Standard Deviation}}{\text{Mean}}\right) \times 100$. The CV is a measure of relative variability, expressing the standard deviation as a percentage of the mean. It is useful for comparing the degree of variation between datasets, even if their means are drastically different.
- **Standard Deviation (SD):** Calculated as the standard deviation of the metric values across the available years. The SD is a measure of absolute variability, indicating the typical distance of the data points from the mean.

These calculations resulted in a new dataset where each row represented a clinic-stratum and included the calculated CV and SD for each metric, along with the clinic's 'LocationAbbr' and the average 'Cycle_Count' for that stratum across the 2020-2022 period ('Avg_Clinic_Volume'). The 'Avg_Clinic_Volume' was calculated as the mean of the reported 'Cycle_Count' values for that clinic-stratum across the years for which metric data was available.

2.5. *Statistical Analysis*

Statistical analyses were conducted to describe the distribution of performance metrics and variability measures and to explore their association with clinic-level factors (Noori et al. 2025; Lara-Cabrera et al. 2025).

2.5.1. *Exploratory Data Analysis*

Prior to variability analysis, descriptive statistics (mean, median, standard deviation, interquartile range,

minimum, maximum) were calculated for each performance and efficiency metric for each year (2020, 2021, 2022), stratified by 'EggSource' and 'AgeGroup' (Otieno et al. 2024). The distribution of 'Cycle_Count' and the geographic distribution of clinics were also summarized.

2.5.2. *Association with Clinic Volume*

The association between the calculated variability measures (CV and SD) and 'Avg_Clinic_Volume' was assessed using Spearman's rank correlation coefficient. Spearman correlation was chosen for its robustness to non-normal distributions and potential non-linear relationships between variability and volume (Stepanov 2024; de Winter et al. 2024). Correlation coefficients (ρ) and associated p-values were reported for each variability measure and stratum.

2.5.3. *Association with Geographic Location*

The association between variability measures and geographic location ('LocationAbbr') was investigated using the Kruskal-Wallis test (Fruchter et al. 2015). This non-parametric test compares the median variability across different states (Haruki et al. 2025). Analyses were limited to states with a minimum number of clinics (e.g., > 5 or > 10, depending on data availability per stratum) to ensure sufficient sample size per group. Test statistics and p-values were reported.

2.5.4. *Multivariable Regression Analysis*

To examine the independent contributions of clinic volume and geographic location to performance variability, Ordinary Least Squares (OLS) linear regression models were developed. Separate models were fitted for each variability measure (CV and SD) within each stratum ('EggSource', 'AgeGroup') as the dependent variable. Independent variables included 'Avg_Clinic_Volume' (treated as a continuous predictor) and 'LocationAbbr' (treated as a categorical predictor, represented by dummy variables with a reference state). Model diagnostics, including linearity, homoscedasticity, and normality of residuals, were examined (Christodoulou et al. 2024). Regression coefficients, standard errors, and p-values were reported for each predictor.

2.6. *Computational Environment*

All data processing and statistical analyses were performed using Python (version 3.9) with standard libraries including pandas (version 1.3.4) for data manipulation, NumPy (version 1.21.5) for numerical operations, SciPy (version 1.7.3) for statistical functions (including Kruskal-Wallis) (Virtanen et al. 2019), and statsmodels (version 0.13.2) for regression modeling. Parallel processing capabilities, where applicable, were utilized using libraries like 'joblib' or 'multiprocessing' to enhance computational efficiency given the size of the dataset and the number of analyses performed. Code and intermediate data files were managed to ensure reproducibility.

3. RESULTS

This study aimed to characterize the year-to-year variability in key U.S. Assisted Reproductive Technology (ART) clinic performance and efficiency metrics between 2020 and 2022, a period marked by the COVID-19 pandemic, and to identify associated clinic-level factors using data from the National ART Surveillance System (NASS).

3.1. *Data cohort and preparation*

The analysis utilized clinic-level aggregate data from the NASS for the reporting years 2020, 2021, and 2022. The initial dataset comprised 1,126,080 records. Key performance and efficiency metrics were identified and extracted based on combinations of descriptive columns. An important limitation encountered during data preparation was the inability to map the metric "Percentage of live births per transfer" (`Perc_LB_Transfer`) from the source data for the specified years and strata, leading to its exclusion from subsequent analysis.

The metrics successfully extracted and analyzed included: Percentage of live births per intended retrieval (`Perc_LB_IntendedRetrieval`), Percentage of live births per actual retrieval (`Perc_LB_ActualRetrieval`), Average number of transfers per intended retrieval (`Avg_Transfers_IntendedRetrieval`), Average number of intended retrievals per live birth (`Avg_IntendedRetrievals_LB`), and Percentage of donor-egg embryo transfer cycles leading to live births (`Donor_Egg_LB_Rate`).

Data were stratified by egg source ('Own' or 'Donor') and patient age group (<35, 35-37, 38-40, >40 for own eggs; 'All Ages Donor' for donor eggs). Clinic volume was represented by the maximum cycle count reported for a given clinic-year-stratum (`Stratum_Cycle_Count`). The distribution of clinic stratum cycle count across the 2020-2022 period is shown in Figure 2, illustrating a highly right-skewed distribution with many low-volume strata and a few high-volume ones. Geographic location was represented by the state abbreviation (`LocationAbbr`). A total of 510 unique clinics were identified across the U.S. states and territories. The geographic distribution of these clinics by state is presented

in Figure 3, showing an uneven concentration in certain states. The data were reshaped into a wide format, with each row representing a unique clinic-year-stratum combination, totaling 6,800 such combinations.

3.2. *Descriptive analysis of ART metrics (2020-2022)*

Descriptive statistics for the key ART metrics were computed for each year, stratified by egg source and age group.

For patients using their own eggs, success rates (`Perc_LB_IntendedRetrieval`, `Perc_LB_ActualRetrieval`) consistently decreased with advancing maternal age across all three years, as expected. For example, in 2020, the mean `Perc_LB_IntendedRetrieval` ranged from 15.9% for the <35 age group to 1.3% for the >40 age group. Similar patterns were observed in 2021. However, the 2022 data showed notable shifts in these percentages for some age groups, with a marked increase in mean `Perc_LB_IntendedRetrieval` for the 35-37 group (27.2%) compared to previous years, while other groups showed smaller changes. The distributions within each stratum, as visualized in boxplots (not shown for all metrics, but illustrative examples of metric distributions are provided later when discussing variability), indicated considerable inter-clinic variability in performance within each year.

Efficiency metrics for own-egg cycles also varied by age. The average number of intended retrievals per live birth (`Avg_IntendedRetrievals_LB`) tended to increase with age, suggesting more cycles are needed for a successful outcome in older patients. The average number of transfers per intended retrieval (`Avg_Transfers_IntendedRetrieval`) was generally below 1.0 across strata, indicating that not all retrieval cycles proceed to transfer. The distribution of this metric across years and age groups is shown in Figure 1.

For donor egg cycles, the mean `Donor_Egg_LB_Rate` was 3.2% in 2020 and 3.9% in 2021 across reporting clinics. A significant data anomaly was observed for 2022, where the `Donor_Egg_LB_Rate` was reported as 0.0% for all 457 clinic-year-stratum instances available in the descriptive analysis. This finding suggests a potential systemic data issue for this specific metric in 2022 within the NASS dataset, rendering direct comparisons and trend analysis for this metric involving 2022 data unreliable.

3.3. *Variability in ART clinic performance (2020-2022)*

Year-to-year variability in each performance and efficiency metric was quantified for each clinic-stratum with

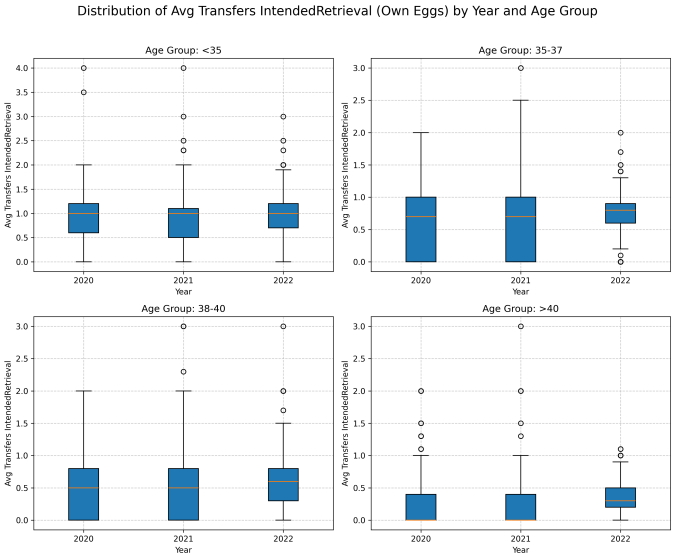

Figure 1. Distribution of average embryo transfers per intended egg retrieval for Own Egg cycles by Year (2020-2022) and Age Group. Boxplots illustrate that this metric is generally below 1.0 across years and age groups, indicating that not all intended retrievals resulted in a transfer, with variation within each group.

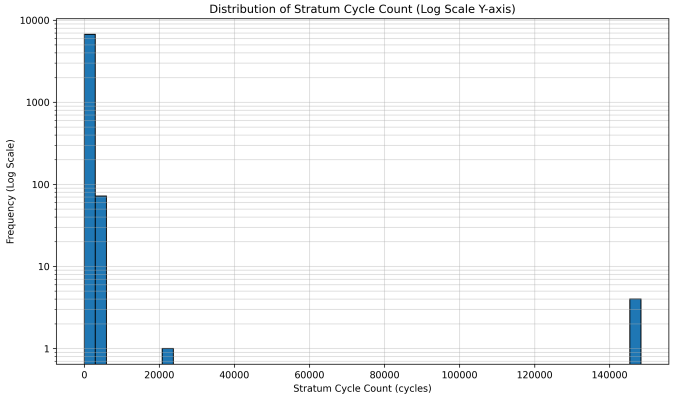

Figure 2. Distribution of clinic stratum cycle count from 2020-2022. The histogram, with a log-scaled frequency axis, shows a highly right-skewed distribution, indicating that the majority of clinic-year-egg source-age group strata have low cycle volumes, while a small proportion have very high volumes.

data in at least two of the three years (2020-2022). Variability was measured using the Coefficient of Variation (CV) and Standard Deviation (SD).

Summary statistics for the calculated CVs and SDs highlighted substantial variability. For own-egg success rates (`Perc_LB_IntendedRetrieval`), the median SD across clinic-strata was 3.70, reflecting the typical absolute fluctuation in percentage points. The median CV was 86.6%, indicating high relative variabil-

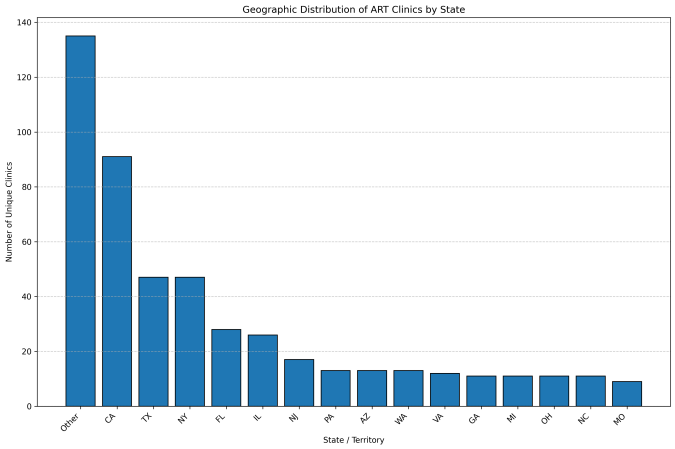

Figure 3. Geographic distribution of unique ART clinics by state/territory (2020-2022). The figure shows the number of clinics per state or territory, indicating an uneven distribution with major hubs in states like California, Texas, and New York.

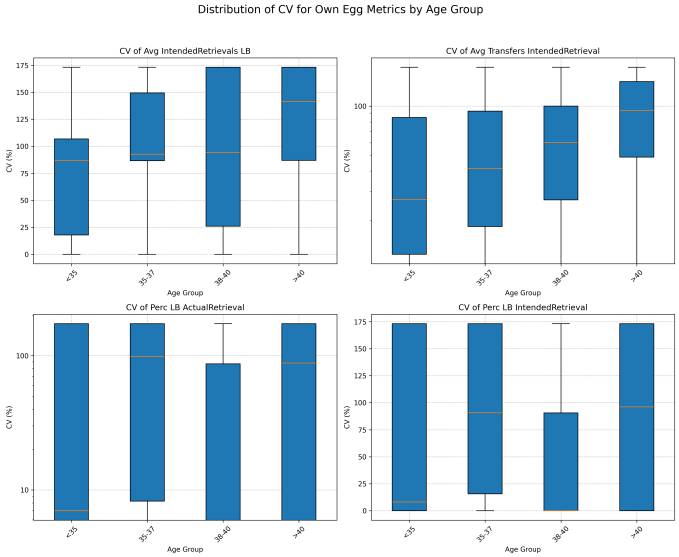

Figure 4. Distribution of clinic-level Coefficient of Variation (CV) for Own Egg ART metrics across age groups. Boxplots show the year-to-year variability for average intended retrievals per live birth, average transfers per intended retrieval, percentage live birth per actual retrieval, and percentage live birth per intended retrieval by age group, indicating differing levels of variability across patient age.

ity, particularly pronounced when mean success rates are low. Similarly, `Perc_LB_ActualRetrieval` showed a median SD of 3.56 and a median CV of 25.1%. For efficiency metrics, `Avg_Transfers_IntendedRetrieval` had a median SD of 0.25 and CV of 53.9%, while `Avg_IntendedRetrievals_LB` showed a median SD of 1.62 and a high median CV of 92.9%.

The distribution of CVs for own-egg ART metrics across age groups is visualized in Figure 4, and the distribution of corresponding SDs is shown in Figure 5. These figures illustrate that the magnitude and distribution of variability differ by metric and age group. CVs for success rates tended to be higher in older age groups, likely due to lower mean success rates in these strata making the CV more sensitive to absolute variations. SDs, representing absolute variability, showed less consistent age-related patterns but were substantial across all age groups.

For the `Donor_Egg_LB_Rate`, the median SD and CV were reported as 0.0%, heavily influenced by the 2022 data anomaly where all reported rates were 0%. However, the mean SD (3.70%) and mean CV (22.96

3.4. *Correlates of performance variability*

Associations between clinic-level variability metrics (CV and SD) and clinic characteristics (`Avg_Clinic_Volume` and `LocationAbbr`) were explored. `Avg_Clinic_Volume` was calculated as the average stratum cycle count for each clinic-stratum across the 2020-2022 period.

3.4.1. *Association with clinic volume*

Spearman's rank correlation coefficient was used to assess the bivariate association between variability measures and `Avg_Clinic_Volume`. For success rate variability in own-egg cycles, both CV (`CV_Perc_LB_IntendedRetrieval`, `CV_Perc_LB_ActualRetrieval`) and SD (`SD_Perc_LB_IntendedRet` `SD_Perc_LB_ActualRetrieval`) showed statistically significant positive correlations with `Avg_Clinic_Volume` across all age groups ($p < 0.001$). For instance, the Spearman Rho for `CV_Perc_LB_IntendedRetrieval` in the <35 age group was 0.436, and for `SD_Perc_LB_ActualRetrieval` in the >40 age group was 0.600. This indicates that clinics with higher average cycle volumes tended to exhibit greater relative and absolute year-to-year variability in their success rates during this period. Examples of these positive correlations are depicted in scatter plots for specific metrics and age groups, such as `CV_Perc_LB_ActualRetrieval` for age <35 (Figure 6) and age 38-40 (Figure 7), and `CV_Perc_LB_IntendedRetrieval` for age 35-37 (Figure 8). Another example for `CV_Perc_LB_ActualRetrieval` in the 35-37 age group is shown in Figure 9.

Conversely, for efficiency metric variability in own-egg cycles, CVs and SDs often showed statistically significant negative correlations with `Avg_Clinic_Volume`.

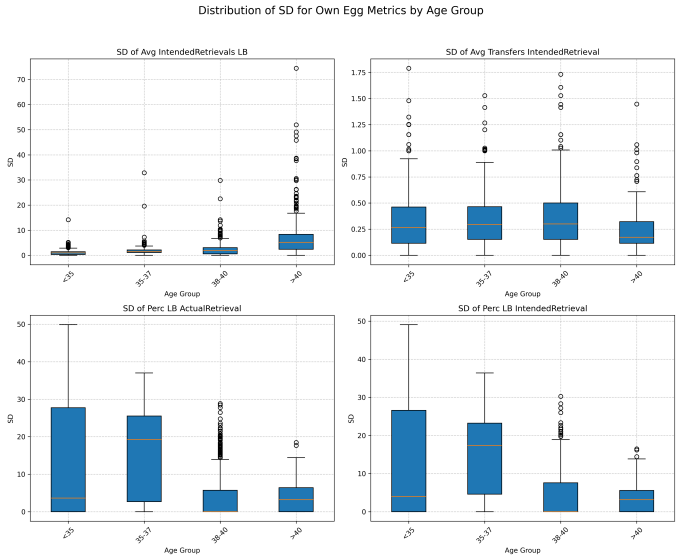

Figure 5. Boxplots showing the distribution of the Standard Deviation (SD) for four U.S. ART clinic performance and efficiency metrics for Own Egg cycles, stratified by patient age group (2020-2022). The metrics are (clockwise from top left): Avg Intended Retrievals per Live Birth, Avg Transfers per Intended Retrieval, Perc Live Birth per Actual Retrieval, and Perc Live Birth per Intended Retrieval. The figure illustrates the year-to-year variability in these metrics across clinics and age groups, showing that SD distributions vary by age, with success rate SDs generally lower in older age groups and efficiency metric SDs showing different patterns.

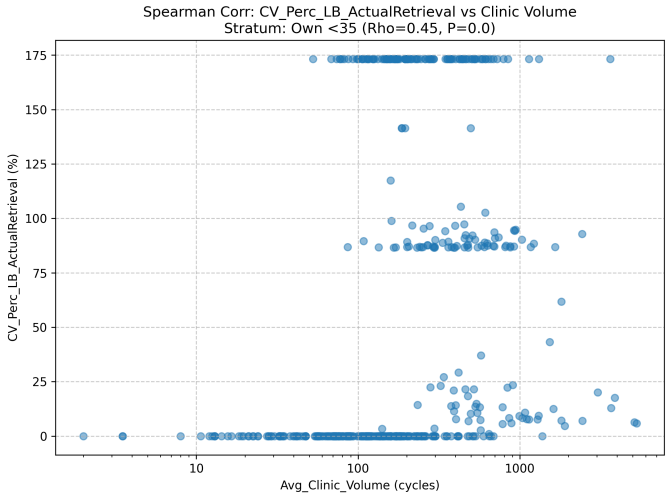

Figure 6. Scatter plot showing the Coefficient of Variation (CV) of the live birth rate per actual retrieval against average clinic volume for own egg cycles in patients aged less than 35 years. The positive correlation (Spearman Rho = 0.45, $p < 0.001$) indicates that higher volume clinics exhibit greater relative variability in this success rate.

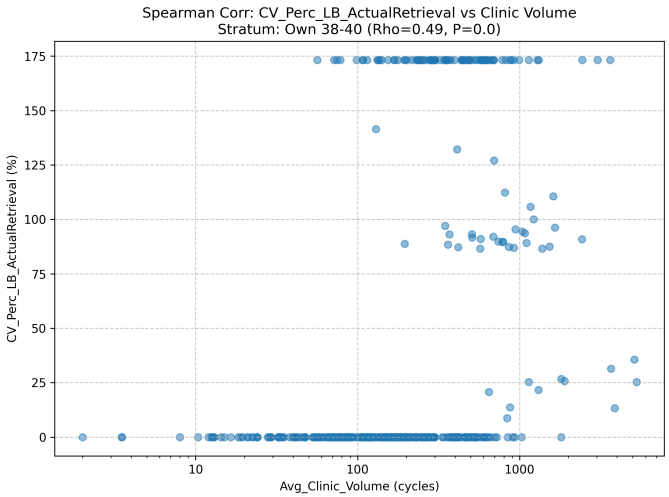

Figure 7. Scatter plot showing the Coefficient of Variation (CV) for the percentage of live births per actual retrieval versus average clinic volume for clinics using patients' own eggs, age 38-40. The positive Spearman correlation ($\rho = 0.49, P < 0.001$) indicates that higher volume clinics are associated with greater relative variability in this success metric.

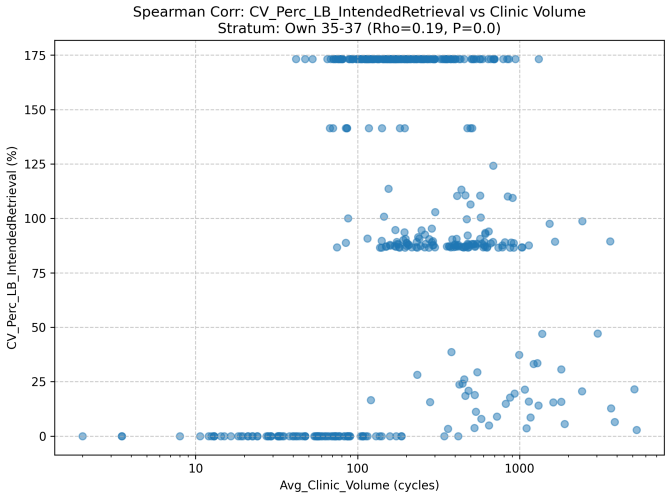

Figure 8. Scatter plot of the coefficient of variation (CV) for percentage live birth per intended retrieval vs. average clinic volume for own egg cycles, age 35-37. The plot indicates that higher volume clinics tend to show greater relative year-to-year variability in this success rate (Spearman Rho=0.19, $P < 0.001$).

For example, `CV_Avg_Transfers_IntendedRetrieval` correlated negatively with volume across most age groups (e.g., Rho=-0.268 for <35), suggesting higher volume clinics experienced less relative variability in the average number of transfers per retrieval. `CV_Avg_IntendedRetrievals_LB` also showed signifi-

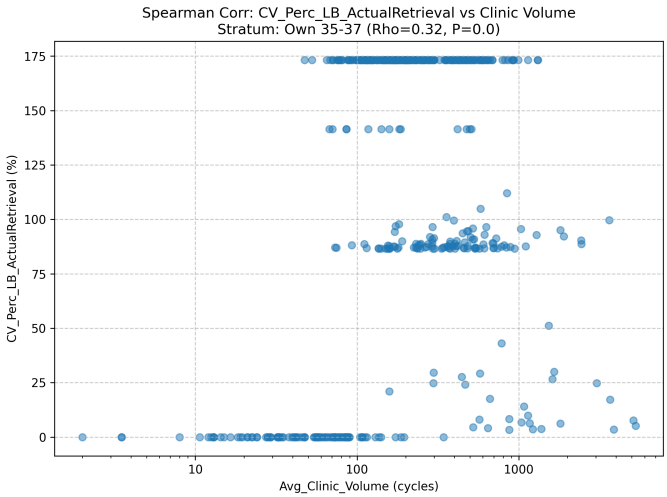

Figure 9. Scatter plot illustrating the positive correlation (Spearman $\rho = 0.32$) between average clinic volume and the relative year-to-year variability (Coefficient of Variation, CV) of the percentage of live births per actual egg retrieval for U.S. ART clinics using Own Eggs in the 35-37 age group. This shows that higher volume clinics tend to have greater relative variability in this success rate. The x-axis is log-scaled.

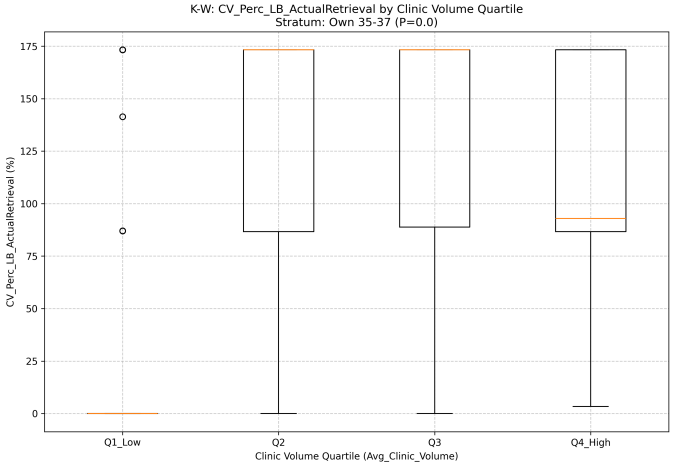

Figure 10. Boxplot showing the distribution of the Coefficient of Variation (CV) for the percentage of actual egg retrieval cycles resulting in live births (`Perc_LB_ActualRetrieval`) among clinics, stratified by average clinic volume quartile, for Own Egg cycles in the 35-37 age group. Clinics in higher volume quartiles generally exhibit greater variability in this success metric.

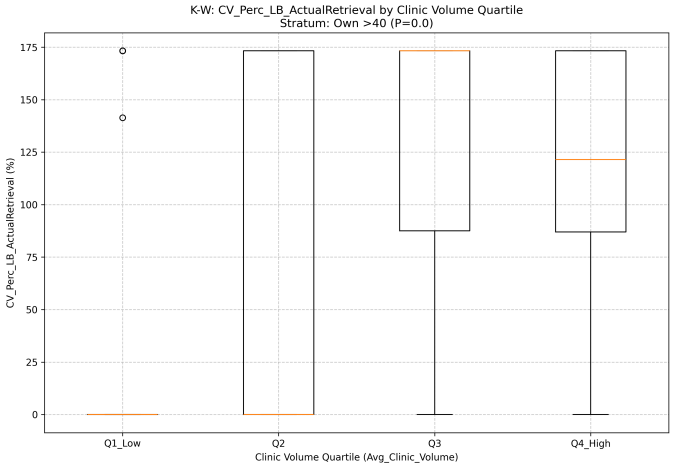

Figure 11. Boxplots showing the distribution of the Coefficient of Variation (CV) for Percentage Live Birth per Actual Retrieval (`CV_Perc_LB_ActualRetrieval`) across clinic volume quartiles for own egg cycles in patients aged >40. Significant differences in CV distribution are observed across quartiles (Kruskal-Wallis $P = 0.0$), indicating that higher volume clinics exhibit greater relative variability in this success rate.

cant negative correlations for younger age groups (<35, 35-37).

Kruskal-Wallis tests comparing variability metrics across clinic volume quartiles largely corroborated these findings, showing statistically significant differences ($p < 0.001$) for most metrics. Clinics in higher volume quartiles were associated with higher success rate variability, as shown for `CV_Perc_LB_ActualRetrieval` for ages 35-37 (Figure 10), >40 (Figure 11), and <35 (Figure 12). Similarly, `CV_Perc_LB_IntendedRetrieval` also showed increasing variability with volume quartile for ages 38-40 (Figure 13) and 35-37 (Figure 15). Conversely, higher volume quartiles were associated with lower efficiency metric variability (for metrics where a negative correlation was observed).

For donor egg cycles, both `CV_Donor_Egg_LB_Rate` and `SD_Donor_Egg_LB_Rate` showed significant positive correlations with `Avg_Clinic_Volume` (Rho=0.418 and 0.429, respectively, $p < 0.001$). Figure 16 visualizes the positive association between CV for donor egg live birth rate and average clinic volume, and Figure 17 shows the distribution of this CV across volume quartiles. These results suggest higher volume was associated with higher variability in donor egg success rates, albeit based on data affected by the 2022 anomaly.

3.4.2. *Association with geographic location (state)*

Kruskal-Wallis tests were conducted to compare median variability metrics across states with sufficient clinic representation (at least 5 clinics per stratum). For most variability metrics and strata, these tests did not reveal widespread statistically significant differences across states. While some isolated instances of significant differences ($p < 0.05$) were observed for specific metric-stratum combinations (e.g., `CV_Perc_LB_IntendedRetrieval` for Own Egg, <35

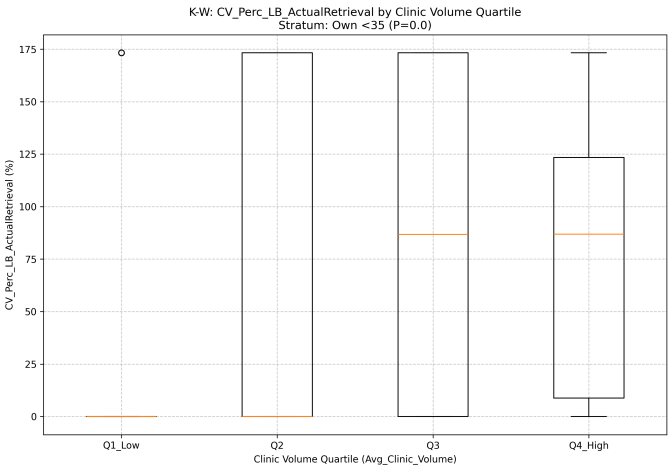

Figure 12. Boxplots showing the distribution of the coefficient of variation (CV) for percentage live birth per actual retrieval for Own Egg cycles in patients <35 years, stratified by clinic volume quartile. The relative year-to-year variability in this success rate is significantly higher in clinics with greater average cycle volume.

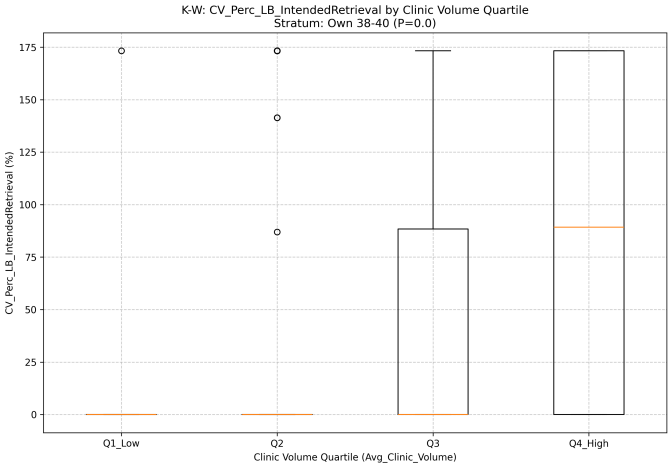

Figure 13. Boxplots showing the distribution of the Coefficient of Variation (CV) for Percentage Live Birth per Intended Retrieval (`Perc_LB_IntendedRetrieval`) for own egg cycles in the 38-40 age group, stratified by average clinic volume quartile. The median CV tends to be higher in clinics with higher average cycle volume (Q3 and Q4_High), indicating greater relative year-to-year variability in this success rate metric for larger clinics. Kruskal-Wallis test indicates a significant difference in CV distribution across volume quartiles ($P < 0.001$).

age group; `CV_Avg_Transfers_IntendedRetrieval` for Own Egg, 35-37 age group), these were not consistent across all metrics or age groups, suggesting state-level differences in variability are less pronounced or uniform compared to the effect of clinic volume.

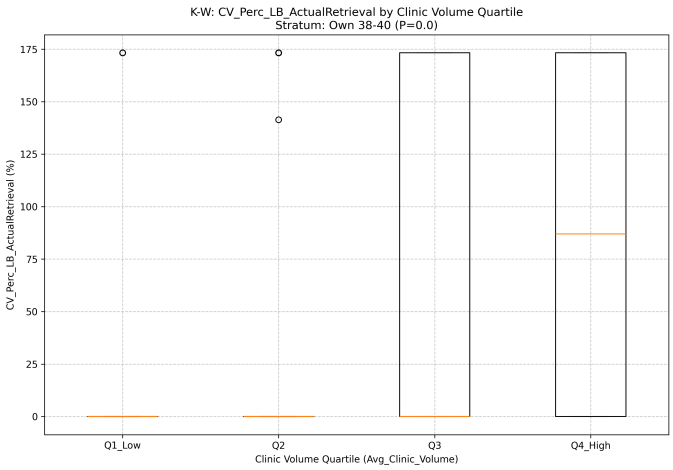

Figure 14. Distribution of the coefficient of variation (CV) for the percentage of live births per actual retrieval (`CV_Perc_LB_ActualRetrieval`) for own egg cycles, age 38-40, by clinic volume quartile. Clinics in higher volume quartiles show greater relative variability in this success rate metric.

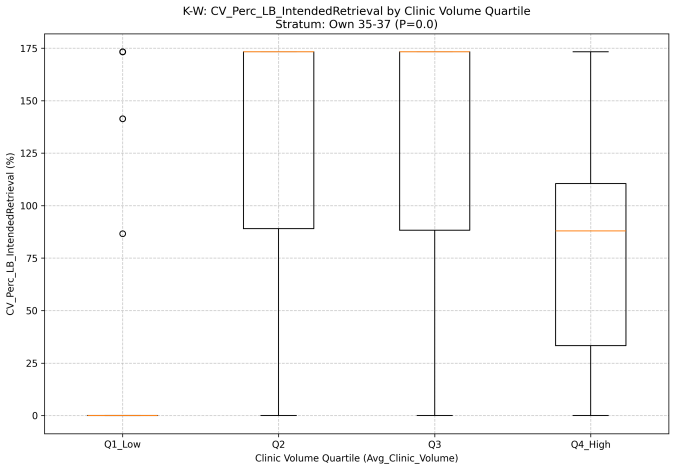

Figure 15. Distribution of the coefficient of variation (CV) for the percentage of intended egg retrievals resulting in live births (Perc_LB_IntendedRetrieval) across clinic volume quartiles for own egg cycles (35-37 age group). The boxplots illustrate that relative year-to-year variability in this success rate metric increases significantly with increasing clinic volume (Kruskal-Wallis $p < 0.001$).

3.4.3. *Multivariable OLS regression analysis*

Ordinary Least Squares (OLS) linear regression models were fitted to examine the combined and independent associations of `Avg_Clinic_Volume` and `LocationAbbr` (state) with each variability metric (CV and SD) within each stratum.

The fitted models generally exhibited low to moderate R-squared values, typically ranging from below 0.10 to

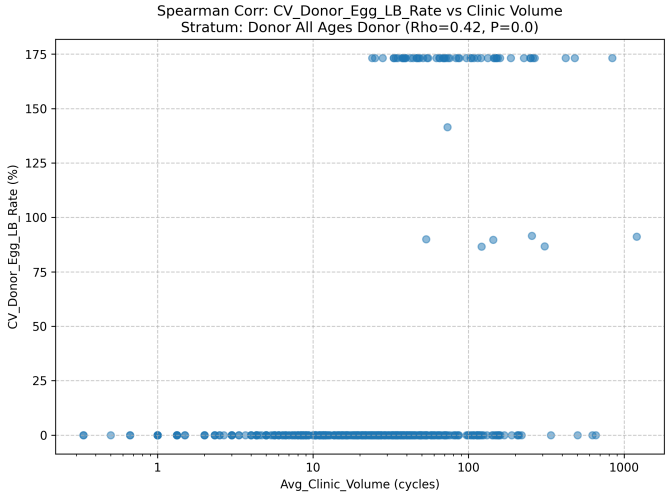

Figure 16. Scatter plot showing the relationship between the Coefficient of Variation (CV) for the Donor Egg Live Birth Rate and average clinic volume for donor egg cycles. Higher average clinic volume is associated with greater relative year-to-year variability in donor egg live birth rates (Spearman $\rho = 0.42$, $p < 0.001$).

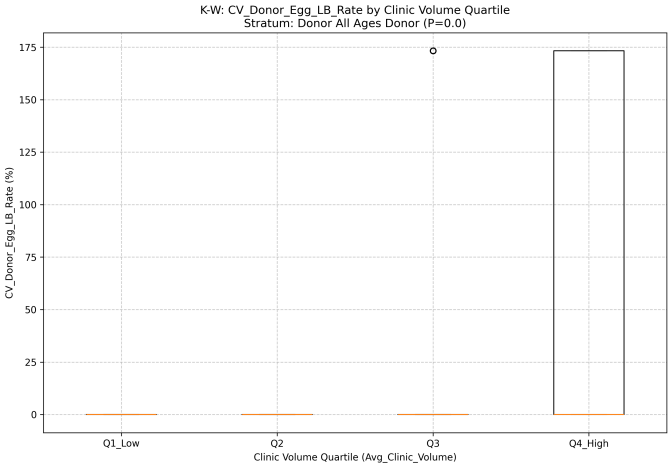

Figure 17. Distribution of the Coefficient of Variation (CV) for the percentage of donor-egg embryo transfer cycles leading to live births (`Donor_Egg_LB_Rate`) by clinic volume quartile for 2020-2022. The figure indicates that clinics with higher average volume exhibit greater relative variability in this success rate.

around 0.19. This indicates that average clinic volume and state, together, explained only a modest proportion of the observed variance in clinic performance variability during the 2020-2022 period.

Consistent with the correlation analysis, `Avg_Clinic_Volume` was a statistically significant predictor in many models. For CVs and SDs of success rates, `Avg_Clinic_Volume` consistently had a positive regression coefficient ($p < 0.05$ or $p < 0.001$ in many strata), confirming that higher average volume was associated with higher success rate variability after adjusting for state. For instance, in the model for `SD_Perc_LB_ActualRetrieval` (Own, >40), the coefficient for `Avg_Clinic_Volume` was 0.0041 ($p < 0.001$). Conversely, for CVs and SDs of efficiency metrics like `Avg_Transfers_IntendedRetrieval`, `Avg_Clinic_Volume` often had a significant negative coefficient (e.g., -0.0235 for `CV_Avg_Transfers_IntendedRetrieval` (Own, 35-37), $p < 0.001$), indicating lower variability in higher volume clinics for these metrics.

After adjusting for clinic volume, some states showed statistically significant differences in variability compared to the reference state (typically California 'CA') for specific metrics and strata. For example, clinics in Nevada (NV) and Virginia (VA) showed significantly higher `CV_Perc_LB_IntendedRetrieval` for the <35 age group compared to CA. Clinics in North Carolina (NC) showed significantly higher `CV_Donor_Egg_LB_Rate` compared to CA. However, many state coefficients were not statistically significant, reinforcing the finding that state-level effects on variability were not uniform or consistently strong across all outcomes and patient groups.

Model diagnostics, including examination of residual plots, indicated that while OLS provided insights into linear associations, some models, particularly those with CV as the dependent variable, showed deviations from normality and potential heteroscedasticity in residuals. Examples of diagnostic plots are shown for models predicting `CV_Perc_LB_ActualRetrieval` for Own Egg age 38-40 (Figure 18), Own Egg age >40 (Figure 19), `CV_Perc_LB_IntendedRetrieval` for Own Egg age >40 (Figure 20), and `CV_Donor_Egg_LB_Rate` for Donor Egg cycles (Figure 21). These deviations suggest that standard assumptions for ordinary least squares regression may not be fully met, and results should be interpreted with consideration for the distributional properties of the variability measures.

3.5. *Data limitations and noteworthy observations*

Several data limitations were encountered during this analysis. The inability to include the important `Perc_LB_Transfer` metric limited the scope of the performance assessment. The significant anomaly in 2022 reporting for `Donor_Egg_LB_Rate`, where a value of 0.0% was recorded for all clinics, severely impacted the analysis of variability for this metric involving the year 2022 and necessitates caution in interpreting these results. The inherent zero-inflation in some success met-

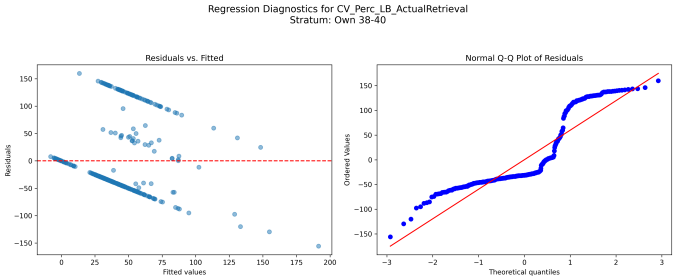

Figure 18. Regression diagnostics for the ordinary least squares model predicting the coefficient of variation for percentage live births per actual retrieval for own egg cycles in patients aged 38-40. The residuals vs. fitted plot and normal Q-Q plot assess model assumptions, revealing potential deviations from linearity, homoscedasticity, and normality of residuals.

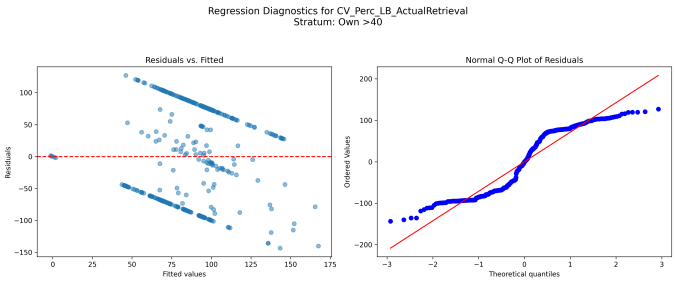

Figure 19. Regression diagnostic plots for the model predicting the Coefficient of Variation (CV) of the percentage of live births per actual retrieval (`CV_Perc_LB_ActualRetrieval`) for Own Egg cycles in the >40 age group. The left panel shows residuals versus fitted values, indicating potential heteroscedasticity. The right panel is a Normal Q-Q plot of residuals, showing deviations from normality, suggesting that standard assumptions for ordinary least squares regression may not be fully met for this model.

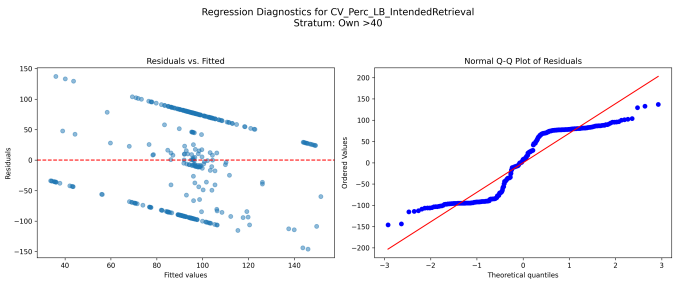

Figure 20. Regression diagnostic plots for the ordinary least squares model predicting the coefficient of variation of the percentage of live births per intended retrieval for patients using their own eggs aged >40. The left panel shows residuals versus fitted values, and the right panel shows the normal Q-Q plot of residuals, illustrating potential deviations from model assumptions.

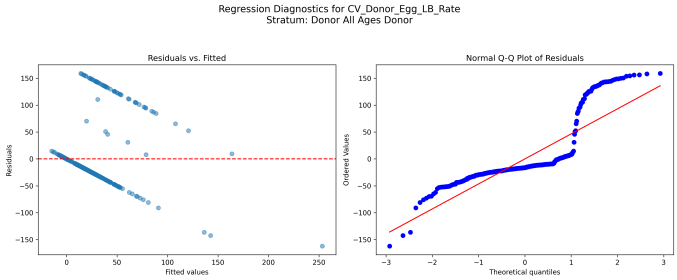

Figure 21. Regression diagnostics for the OLS model predicting the Coefficient of Variation of the Donor Egg Live Birth Rate (`CV_Donor_Egg_LB_Rate`) for donor egg cycles (All Ages Donor stratum). The plots show deviations from linear model assumptions, including potential heteroscedasticity (residuals vs. fitted) and non-normality of residuals (Normal Q-Q plot), suggesting limitations in the model fit for this metric, particularly in the context of the 2022 data anomaly.

rics, particularly for older age groups or smaller clinics, can influence variability calculations, making CVs especially sensitive to low mean values. Furthermore, while the study period coincided with the COVID-19 pandemic, the NASS dataset does not contain direct measures of pandemic impact on clinic operations, limiting the ability to directly attribute observed variability patterns to specific pandemic-related disruptions. Finally, the definition of clinic volume used, based on stratum cycle counts, serves as a reasonable proxy but may have nuances depending on how cycle counts are reported in the source data.

In summary, the analysis revealed substantial year-to-year variability in U.S. ART clinic performance and efficiency metrics between 2020 and 2022. Contrary to a simple assumption that larger volume might buffer against variability, higher average clinic volume was associated with *higher* relative and absolute variability in success rates for both own-egg and donor-egg cycles, while simultaneously being associated with *lower* variability in some efficiency metrics. Geographic location showed some stratum-specific associations with variability, but these were not as consistent or widespread as the associations with clinic volume. Clinic volume and state together explained only a modest proportion of the observed variance in variability. These findings highlight the complex nature of ART clinic performance fluctuations during a period of potential external stress and underscore the importance of acknowledging data reporting challenges in national surveillance systems.

4. CONCLUSIONS

This study aimed to characterize the year-to-year variability in key U.S. Assisted Reproductive Technology

(ART) clinic performance and efficiency metrics during the COVID-19 pandemic years (2020-2022) and explore associations with clinic volume and geographic location using data from the National ART Surveillance System (NASS). Understanding this variability is crucial for patients, clinics, and public health surveillance, especially during periods of potential disruption.

Utilizing clinic-level aggregate data from NASS for 2020, 2021, and 2022, we quantified year-to-year variability for several success and efficiency metrics (percentage live birth per intended/actual retrieval, average transfers per intended retrieval, average intended retrievals per live birth, and donor egg live birth rate) using the Coefficient of Variation (CV) and Standard Deviation (SD). Analysis was stratified by egg source (own vs. donor) and patient age group. We then investigated the association between these variability measures and average clinic volume (cycle count) and geographic location (state) using Spearman correlation and Ordinary Least Squares regression. Significant data limitations were noted, including the exclusion of the important percentage live birth per transfer metric due to data availability issues and a significant anomaly in the 2022 reporting of the donor egg live birth rate, which showed 0.0% for all clinics.

The analysis revealed substantial year-to-year variability in ART clinic performance and efficiency metrics across the 2020-2022 period. Both relative (CV) and absolute (SD) variability were considerable for many metrics and patient strata. Counterintuitively, higher average clinic volume was consistently associated with *higher* relative and absolute variability in success rates for both own-egg and donor-egg cycles. Conversely, higher average clinic volume was associated with *lower* variability in some efficiency metrics, such as the average number of transfers per intended retrieval. Geographic location demonstrated some state-specific associations with variability, but these were not uniform across all metrics or patient groups and were less consistent than the associations with clinic volume. Overall, average clinic volume and state location explained only a modest proportion (low to moderate R-squared values) of the observed variance in performance variability.

From these findings, we have learned several key points regarding ART clinic performance during the pandemic era. First, significant fluctuations in clinic-level success and efficiency metrics occurred from year to year, highlighting that performance is not static, particularly under external stressors. Second, the relationship between clinic volume and performance variability is complex and metric-dependent; larger clinics, while potentially offering advantages in scale and resources, experienced greater swings in success rates compared to smaller clinics during this period. This suggests that higher volume may not simply buffer against variability in success outcomes and could potentially be linked to factors such as managing a larger diversity of cases, greater complexity of operations, or the absolute number of events driving larger standard deviations. In contrast, larger volume clinics appeared to maintain more stable efficiency metrics, possibly reflecting more standardized operational processes. Third, while geographic factors may play a role in some instances, they do not appear to be a dominant or consistent driver of performance variability across the board. Finally, the study underscores the critical importance of data quality and consistency in national surveillance systems like NASS for accurate characterization and analysis of trends and variability in ART outcomes. Data anomalies and limitations in reported metrics can significantly impact the ability to draw robust conclusions.

In conclusion, this study provides novel insights into the dynamics of U.S. ART clinic performance variability during a period marked by the COVID-19 pandemic. The findings challenge simple assumptions about clinic volume as a universal buffer against variability, revealing differential effects on success versus efficiency metrics. They emphasize the need to consider multiple clinic-level factors and acknowledge data limitations when interpreting ART outcomes reported through national surveillance. Further research is needed to explore other potential drivers of variability, including specific clinic operational characteristics, adaptations implemented during the pandemic, and the underlying patient populations served, to better inform quality improvement efforts and support prospective patients.

REFERENCES

Abbasian, M., Khatibi, E., Azimi, I., et al. 2024, Foundation Metrics for Evaluating Effectiveness of Healthcare Conversations Powered by Generative AI. https://arxiv.org/abs/2309.12444

Bandara, K., Bergmeir, C., Campbell, S., Scott, D., & Lubman, D. 2020, Towards Accurate Predictions and Causal 'What-if' Analyses for Planning and Policy-making: A Case Study in Emergency Medical Services Demand. https://arxiv.org/abs/2004.12092

Bedi, S., Cui, H., Fuentes, M., et al. 2025, MedHELM: Holistic Evaluation of Large Language Models for Medical Tasks. https://arxiv.org/abs/2505.23802

Bendinelli, T., Dox, A., & Holz, C. 2025, Exploring LLM Agents for Cleaning Tabular Machine Learning Datasets. https://arxiv.org/abs/2503.06664

Caldera, L., Masci, C., Cappozzo, A., et al. 2024, Uncover mortality patterns and hospital effects in COVID-19 heart failure patients: a novel Multilevel logistic cluster-weighted modeling approach. https://arxiv.org/abs/2405.11239

Chen, J., Wei, Z., Zhang, W., Hu, Y., & Zhang, Q. 2025, CliniChat: A Multi-Source Knowledge-Driven Framework for Clinical Interview Dialogue Reconstruction and Evaluation. https://arxiv.org/abs/2504.10418

Christodoulou, E., Reinke, A., Houhou, R., et al. 2024, Confidence intervals uncovered: Are we ready for real-world medical imaging AI? https://arxiv.org/abs/2409.17763

Dai, W., Chen, P., Lu, M., et al. 2025, Data Foundations for Large Scale Multimodal Clinical Foundation Models. https://arxiv.org/abs/2503.07667

de Winter, J. C. F., Gosling, S. D., & Potter, J. 2024, Comparing the Pearson and Spearman Correlation Coefficients Across Distributions and Sample Sizes: A Tutorial Using Simulations and Empirical Data, doi: https://doi.org/10.1037/met0000079

Fruchter, N., Miao, H., Stevenson, S., & Balebako, R. 2015, Variations in Tracking in Relation to Geographic Location. https://arxiv.org/abs/1506.04103

Goyle, K., Xie, Q., & Goyle, V. 2023, DataAssist: A Machine Learning Approach to Data Cleaning and Preparation. https://arxiv.org/abs/2307.07119

Guan, S., Samala, R. K., & Chen, W. 2022, Informing selection of performance metrics for medical image segmentation evaluation using configurable synthetic errors. https://arxiv.org/abs/2212.14828

Harris, T., Jayasundara, P., Ragonnet, R., et al. 2024, Apparent structural changes in contact patterns during COVID-19 were driven by survey design and long-term demographic trends. https://arxiv.org/abs/2406.01639

Haruki, Y., Kato, K., Enami, Y., et al. 2025, Development of Automated Data Quality Assessment and Evaluation Indices by Analytical Experience. https://arxiv.org/abs/2504.02663

Irons, N. J., & Raftery, A. E. 2024, US COVID-19 school closure was not cost-effective, but other measures were. https://arxiv.org/abs/2411.12016

Jones, P., Liu, W., Huang, I.-C., & Huang, X. 2025, Examining Imbalance Effects on Performance and Demographic Fairness of Clinical Language Models. https://arxiv.org/abs/2412.17803

Lara-Cabrera, R., Gonzalez-Pardo, A., & Camacho, D. 2025, Statistical Analysis of Risk Assessment Factors and Metrics to Evaluate Radicalisation in Twitter, doi: https://doi.org/10.1016/j.future.2017.10.046

Lee, G. Y., Alzamil, L., Doskenov, B., & Termehchy, A. 2021, A Survey on Data Cleaning Methods for Improved Machine Learning Model Performance. https://arxiv.org/abs/2109.07127

Montoya, L. M., Geng, E. H., Adhiambo, H. F., & Petersen, M. L. 2025, Estimation and Evaluation of the Resource-Constrained Optimal Dynamic Treatment Rule: An Application to HIV Care Retention. https://arxiv.org/abs/2502.14763

Noori, M., Valiante, E., Vaerenbergh, T. V., Mohseni, M., & Rozada, I. 2025, A Statistical Analysis for Per-Instance Evaluation of Stochastic Optimizers: How Many Repeats Are Enough? https://arxiv.org/abs/2503.16589

Oganisian, A., Hogan, J., Sang, E., et al. 2024, Bayesian Counterfactual Prediction Models for HIV Care Retention with Incomplete Outcome and Covariate Information. https://arxiv.org/abs/2410.22481

Otieno, D. O., Abri, F., Siami-Namini, S., & Namin, A. S. 2024, The Accuracy of Domain Specific and Descriptive Analysis Generated by Large Language Models. https://arxiv.org/abs/2405.19578

Oveson, A., Girvan, M., & Gumel, A. 2025, Modeling the impact of hospitalization-induced behavioral changes on SARS-COV-2 spread in New York City. https://arxiv.org/abs/2501.06941

Parker, F., Ganjkhanloo, F., Martínez, D. A., & Ghobadi, K. 2024, Optimal Hospital Capacity Management During Demand Surges. https://arxiv.org/abs/2403.15738

Silver, D. H., Feder, M., Gold-Zamir, Y., et al. 2020, Data-Driven Prediction of Embryo Implantation Probability Using IVF Time-lapse Imaging. https://arxiv.org/abs/2006.01035

Stepanov, A. 2024, On Correlation Coefficients. https://arxiv.org/abs/2405.16469

Tran, T. N., Adler, T., Yamlahi, A., et al. 2022, Sources of performance variability in deep learning-based polyp detection. https://arxiv.org/abs/2211.09708

Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2019, SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python, doi: https://doi.org/10.1038/s41592-019-0686-2