PharmaSUG 2025 - Paper ST267

Handling Missing Data in External Control Arms: Best Practices, Recommendations, and SAS Code Examples

Yutong Zhang, LLX Solutions, LLC

ABSTRACT

External control arm (ECA) studies provide an alternative approach when traditional randomized controlled trials (RCTs) are not feasible. However, missing data in ECA are facing a critical challenge due to heterogeneous data sources, unstructured visit schedules, and the absence of randomization. These factors increase the likelihood of Missing Not at Random (MNAR), which can undermine the robustness of treatment effect estimates if not properly addressed.

This paper provides a practical review of missing data handling strategies in ECAs, including guidance on selecting appropriate methods based on the missingness mechanism. We discuss mixed models for repeated measures (MMRM), while emphasizing the flexibility and regulatory acceptance of multiple imputation (MI) under the Missing at Random (MAR) assumption. We also introduce sensitivity analysis techniques under MNAR, that are essential for evaluating the robustness of findings.

A two-step framework is recommended: perform the primary analysis under MAR, followed by sensitivity analyses to assess the impact of potential MNAR scenarios. To support implementation, SAS code examples are provided for each method.

INTRODUCTION

ECA studies have gained growing attention in recent years as an alternative to traditional randomized controlled trials, especially in rare disease settings. By using existing data, such as from electronic health records (EHRs), patient registries, or historical trials, ECAs offer a way to evaluate treatment effects more efficiently and in real-world contexts. This approach is particularly valuable in oncology, where the need to avoid placebo or non-treatment arms can make RCTs infeasible.

However, the increased reliance on retrospective and real-world data introduces unique challenges around missing data. ECAs often rely on datasets that were not collected for research purposes. As a result, the structure, timing, and completeness of the data vary widely. This could cause loss of statistical power, and the violation of assumptions required for imputation.

Given these complexities, missing data in ECA studies cannot be addressed with a regular solution. It's important to design a case-by-case missing data imputation method among different ECAs, or even among different data parts in a single ECA. This paper focuses on guiding the selection of imputation strategies tailored to different missingness mechanisms, with practical examples and regulatory considerations in mind. All imputation methods were performed using SAS software version 9.4.

CHALLENGES OF MISSING DATA IN EXTERNAL CONTROL ARMS

Missing data in ECA studies presents several unique challenges. These challenges are primarily driven by the nature of real-world data and the absence of randomization, which complicates both data quality and the assumptions required for valid statistical inference.

DATA HETEROGENEITY ACROSS SOURCES

ECA studies often draw data from heterogeneous sources, such as EHRs, insurance claims, registries, or historical clinical trials. These data sources vary in structure, completeness, and purpose. For example, EHRs are primarily designed for clinical care rather than research, and may lack standardized fields or consistent coding. As a result, key variables may be systematically missing or inconsistently recorded across sources.

HIGH RATES OF MISSINGNESS IN KEY COVARIATES

Due to the variability in data collection practices, critical variables needed for propensity score adjustment, subgroup analyses, or outcome modeling may be missing in a large portion of the external dataset. This may weaken the robustness of analysis between the treatment and control groups.

UNSTRUCTURED VISITS

Unlike RCTs, which follow protocol-specified visit schedules, data collection in ECAs is far less controlled. This results in inconsistent outcome measurement intervals, which can complicate the application of standard longitudinal analysis methods and make certain imputation approaches (e.g., MMRM or LOCF) less appropriate.

MISSING DATA MECHANISMS

Before selecting an imputation method, understanding the missing data mechanism is essential. The three primary mechanisms are:

MISSING COMPLETELY AT RANDOM (MCAR)

MCAR refers to a situation where data is missing randomly, independent of observed and unobserved variables. Under the MCAR assumption, the analysis set is a random sample of the full dataset, and any statistical analysis performed on this set will produce unbiased estimates. However, MCAR is a rare assumption, especially in ECA.

MISSING AT RANDOM (MAR)

Missingness is related to observed variables but not to the value of the missing data itself. This means that the missing values can be explained or predicted by observed values. MAR is a more flexible and realistic assumption than MCAR, and many modern imputation methods are built under the MAR framework. Multiple imputation (MI), predictive mean matching (PMM), and regression-based imputation can provide reliable estimates for MAR.

In ECA, assuming MAR could be reasonable and practicable when handling missing data, especially when demographics and clinical characteristics are well collected.

MISSING NOT AT RANDOM (MNAR)

Missing Not at Random (MNAR) refers to situations where the probability of a data point is missing depends on the unobserved value itself. This makes MNAR the most challenging type of missingness to handle because the missingness cannot be fully explained using observed data alone.

A common example in clinical studies is a patient who deteriorates rapidly due to disease progression. This patient might miss follow-up visits, drop out of the healthcare system, or even pass away — resulting in missing outcome data that is directly related to their poor health status. Since this information is unobserved, standard imputation methods that MAR cannot correct the bias introduced by the missingness. As a result, MNAR often occurs on key endpoints, which raises a serious concern in the robustness and consistency of statistics hypothesis.

The goal of handling MNAR is not to find the true values but to understand how much the missing data could affect the results. When handling missing data in clinical trials, it's common practice to assume data are MAR for the primary analysis and to conduct sensitivity analyses under the MNAR assumption. This approach aligns with FDA guidance (U.S. Food and Drug Administration, 2023) and is widely adopted in the industry.

The choice of imputation method depends heavily on the underlying missingness mechanism, which we address in the following two sections.

IMPUTATION METHODS

The following methods are commonly used when data is assumed to be MCAR or MAR

IMPUTATION UNDER MCAR ASSUMPTION

Although Missing Completely at Random (MCAR) is rarely satisfied in practice, a few simple imputation methods are based on this assumption. **The complete case analysis**, the most straightforward approach, excludes any records with missing data but can lead to biased results and reduced power if MCAR does not hold. **Mean or median imputation** replaces missing values with a constant average, distorting the distribution and underestimating variance. In longitudinal settings, **Last Observation Carried Forward (LOCF)** carries forward the most recent observed value, assuming stability over time — an assumption that often results in biased treatment estimates and underestimated variability. For longitudinal data with scheduled visits and a small number of intermediate missing values, **linear interpolation** may also be used. However, it can only be applied when missing values occur between two observed points, not at the beginning or end. Due to these limitations and the rarity of true MCAR, these methods are generally not recommended for use in inferential analyses.

MIXED MODELS FOR REPEATED MEASURES (MMRM)

Mixed Models for Repeated Measures (MMRM) are commonly used in randomized controlled trials to handle longitudinal missing data under the MAR assumption. MMRM leverages the correlation structure in repeated measurements and accounts for within-subject variability over time. While this method is well-established in trial settings, its application in ECAs is limited. Most ECA datasets, particularly those derived from electronic health records (EHRs) or registries, lack structured visits. This violates the assumptions of MMRM, which expects consistent time points across individuals. However, in cases where the external data has a well-aligned visit schedule or is drawn from a historical clinical trial with protocolized assessments, MMRM may still be applicable. Using **PROC MI** in SAS (Program 1).

MULTIPLE IMPUTATION (MI)

Unlike single imputation, MI acknowledges that there is more than one plausible value for each missing point. The primary thought of MI is to create multiple datasets by estimating missing values based on observed data and then pool the results to maintain variability. The core assumption for MI is MAR, violation may cause bias.

There are three steps for MI analysis:

Step 1: Imputation

MI creates m versions of the dataset for each missing value, using a model-based imputation or predictive mean matching (PMM) method. For a better estimate, recommend m>=20. Using **PROC MI** in SAS (Program 2).

PMM is a method used within the first step in MI. For each missing value, PMM finds observed cases with similar predicted values and then randomly selects one of their actual values to impute (Program 3). Compared with model-based imputation, PMM is more robust to non-normal distribution, making it ideal for skewed or bounded variables.

Step 2: Analysis

Each imputed dataset is analyzed separately using the model of interest (e.g., logistic regression, survival analysis).

Step 3: Pooling

Combine the results using Rubin's Rules to get P-value and confidence interval. Using **PROC MIANALYZE** in SAS (Program 4).

SENSITIVITY ANALYSIS

When missingness is believed to be MNAR, standard imputation methods may yield biased results. This section introduces sensitivity analyses and modeling strategies specifically designed to assess and address MNAR scenarios.

Sensitivity analysis is used to test how the results would change under different assumptions about the missing data mechanism, especially when the correctness of the MAR assumption is difficult to determine.

DELTA ADJUSTMENT

Delta adjustment is the most widely used and accessible version in Pattern Mixture Models (PMMs). The idea is to apply a systematic shift (delta) to the imputed values to simulate what would happen if missing values were systematically worse than the observed data. The goal is to test whether the primary result remains consistent when subject to such plausible deviations.

Delta adjustment follows the same steps as MI, with a simple post-imputation adjustment (apply a fixed shift to the imputed values in imputed datasets from PROC MI) before analysis.

TIPPING POINT ANALYSIS

Tipping point analysis builds on the delta adjustment concept by applying a range of delta values to determine the exact point at which the study conclusion changes — typically when a treatment effect loses statistical significance. A p-value–delta plot visualizes how the statistical changes as increasing shifts are applied to the imputed values. This approach helps identify how sensitive the findings are to the assumptions about missing data.

To determine whether a tipping point is realistic, the identified shift should be compared to the distribution of observed data. If the tipping point is substantially larger than the observed differences between groups or falls outside the typical clinical range, the result is considered robust.

SUMMARY

Delta adjustment and tipping point analysis are closely related sensitivity analysis techniques used to assess the impact of MNAR missing data. In practice, both involve shifting imputed values by a fixed amount, but they differ in interpretation:

- Delta adjustment tests whether the study result remains robust under specific, clinically plausible shifts in missing data values.
- Tipping point analysis systematically increases the delta to identify the threshold at which the study conclusion changes and then evaluates whether that tipping point is likely to occur in reality.

Both approaches support transparent reporting and are widely accepted by regulatory agencies.

RECOMMENDED IMPUTATION STRATEGY

To address the special challenges in ECA, the FDA recommends assuming MAR for the primary analysis, using established methods such as multiple imputation or mixed models, and then performing sensitivity analyses under MNAR assumptions. In other words, sensitivity analysis is not optional in ECA. This two-step strategy ensures that study conclusions remain robust even if the MAR assumption does not fully hold.

STEP 1: ASSESS MISSING DATA PATTERNS

Determine if data is MCAR or MAR using exploratory analysis techniques. Here are several methods and tests to be referred:

1. Visualize Missing Data: Generate a missingness heatmap and correlation plot to check whether certain variables are related in missingness.
2. Little’s MCAR Test: Compares means and covariances of observed data across missing vs. non-missing groups using maximum likelihood estimation (MLE).

This step could be unnecessary if assuming MAR in study design.

STEP 2: SELECT AN IMPUTATION METHOD

Match the imputation technique with the missing data mechanism to minimize bias.

1. MCAR: Complete Case Analysis, Mean/Median Imputation, LOCF
2. MAR: MMRM, MI

STEP 3: SENSITIVITY ANALYSIS TO ASSESS ROBUSTNESS

Regulators require a thorough sensitivity analysis to demonstrate that study conclusions remain robust under different imputation scenarios.

SAS CODE EXAMPLES

```
proc mixed data=mydata;
class trt01p avisitn subjid;
model chg = trt01p avisitn trt01p*avisitn base base*avisitn/solution;
repeated avisitn/ type=un subject=subjid;
ods output lsmestimates=lsm estimates=lsest convergencestatus=conv;
run;
```

Program 1. SAS Code for MMRM

```
proc mi data= mydata nimpute=100 out=mi seed=54321;
class trt01p;
var age ecog response;
run;
```

Program 2. SAS Code for MI

```
proc mi data= mydata nimpute=100 out=mi;
class trt01p;
fcs regpmm(age ecog response);
var age ecog response;
run;
```

Program 3. SAS Code for PMM in MI

```
proc mianalyze data=reg_rst;
class trt01p;
modeleffects lsmean;
stderr stderr;
ods output parameterestimates=params;
run;
```

Program 4. Sas Code for Pooled Step in MI

CONCLUSION

Handling missing data in ECA studies is rarely straightforward. Given the variability in data sources, visit schedules, and underlying assumptions, a case-by-case approach is essential. The two-step framework—conducting a primary analysis under the MAR assumption, followed by sensitivity analyses under MNAR—offers a practical and regulator-aligned strategy for ensuring robustness. This approach aligns with FDA expectations and provides a solid foundation for ensuring the validity of treatment effect estimates under ECAs.

REFERENCES

U.S. Food and Drug Administration. (2023). *Considerations for the design and conduct of externally controlled trials for drug and biological products: Draft guidance for industry.* U.S. Department of Health and Human Services.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Yutong Zhang
Yutong.zhang@llxsolution.com