PSI

# Timepoint selection for long-term PRO data modelling in oncology trials

**Patient Reported Outcomes Session**
**PSI Conference 2026**
**Belfast, 15-17th June 2026**

Rigazio A, Skaltsa K, Kral P, Pestana C and Bell J

# Agenda

- Background and objectives
- Data Generation
- Methods
  - Analysis model
  - Performance measures
- Results
  - Results: 5 visits – scenarios 1A+1B
  - Results: 15 visits – scenarios 1A+1B
  - Results: 25 visits – scenarios 1A+1B
- Key Findings, Limitations and Next Steps

# Background and Objectives

- MMRM is widely used for longitudinal PRO data. Researchers are frequently interested in **long-term treatment effects** to communicate with clinicians and payers, requiring the inclusion of several timepoints.
- In oncology trials, substantial intercurrent events with long and variable post-treatment follow-up PRO collection schedules may lead to **increasing missingness**.
  - › Under these conditions, MMRM estimates may become biased, unstable, or sensitive to missing data patterns.

## How many timepoints should we include in the MMRM?

- The number of timepoints to include in the MMRM becomes a **critical modelling decision**, particularly when a primary clinically meaningful timepoint is difficult to define.
- Current practice relies on **arbitrary thresholds** for data availability (e.g., 10%, 30%, 60%, N=10), lacking consistency and limited interpretability.

## Objective of this work

- Establish **evidence-based thresholds** for timepoint inclusion to improve reliability and interpretability of MMRM analyses in longitudinally collected PRO data in oncology
- This is achieved by means of a **simulation study** that assessed how model performance is affected by drop-out rates, differential attrition between arms, number of visits, sample size and missingness mechanism

MMRM = Mixed Model for Repeated Measures; PRO = Patient-reported Outcome.

# Simulating Complete and Incomplete Data

- Study design: Two-arm randomized (Active vs Control) with 300 patients per arm and repeated PRO assessments
- Outcome: change from baseline (**0–100 scale**) simulated using **truncated multivariate normal distribution**
  - Timepoint-specific *mean and standard deviation* selected from reference oncology trials
  - Correlation structure: *heterogeneous Toeplitz (TOEPH),* reflecting decreasing correlations over time and time-specific within-subject variances
- Simluation parameters and attrition rates informed by a review of several phase III clinical oncology trials
- **10 *simulated complete datasets*** (**SCDs**) for each combination of the following parameters:

| Parameter | Values |
|---|---|
| **Number of timepoints (including baseline)** | **5** *(~Week 12-24*)*, **15** *(~Week 42-84*)*, **25** *(~Week 72-144*)* |
| **Treatment Effect Size** (Cohen's d) | **0.2** *(small)*, **0.5** *(moderate)*, **0.8** *(large)* |

- **100 *simulated incomplete datasets*** (**SIDs**) generated for each SCD with **monotone dropout under MAR**** under the 4 scenarios below:

| | SID Scenario 1: Equal attrition | | SID Scenario 2: Unequal attrition | |
|---|---|---|---|---|
| | **1A** | **1B** | **2A** | **2B** |
| **Dropout Rate by Timepoint** | • **5%** in both arms | • **10%** in both arms | • **2%** in Active<br>• **5%** in Control | • **5%** in Active<br>• **10%** in Control |

* PRO assessments assumed every 3-6 weeks.
** Dropout applied probabilistically at each timepoint conditional on observed prior outcomes.

# Analysis Model

- **For each simulated dataset,** the following model was applied:

  Change from baseline = baseline + treatment + timepoint+ (baseline*timepoint) + (treatment*timepoint)

  - Parameters estimated using REML (Newton-Raphson algorithm)
  - **Unstructured** covariance matrix to model within-patient correlations
  - Least square (LS) means difference between arms at each timepoint + standard errors (SEs)
- **Pooling Estimates into each Simulated Scenario:**
  - LS mean differences and SEs combined using Rubin's rules (not presented here)

For this exercise, we assume the MMRM is the desired model to analyze the data. Examining the plausibility of this assumption is not in scope.

MAR = Missing At Random; MMRM = Mixed Model for Repeated Measures; REML=restricted maximum likelihood.

# Performance Measures

## Bias

Difference between SID and SCD LS mean treatment difference estimates

$$Bias_{ij} = \Theta_{ti} - \widehat{\Theta}_{tij}$$

- $\Theta_{ti}$ being the i$^{th}$ (i=1,…,10) **SCD** (LS mean) **"true" difference** at timepoint t
- $\widehat{\Theta}_{tij}$ being the j$^{th}$ (j=1,…100) **SID** (LS mean) estimate within the $i$-th SCD at timepoint t

Distribution of bias assessed using box plots
Proportions of deviations from true difference by magnitude using bar charts

## Interpretability - Cohen's categories

Difference between **SID** and **SCD** in Cohen's d categories of treatment difference:

< 0.2 (*negligible*) | 0.2- < 0.5 (*small*) | 0.5 - < 0.8 (*moderate*) | ≥ 0.8 (*large*)

$$Cohen's\ d = LS\ Mean\ Difference/Model, based\ SD_{Pooled}$$

Proportions of misclassifications in **"true" Cohen's d category** using bar charts

## Coverage - 95% Confidence Interval (CI)

Proportion of **SID** simulations where the 95% CI includes the **"true" difference**

$$Coverage_t(\%) = \frac{1}{N_{SCD} * N_{SID}} \sum_{i=1}^{10} \sum_{j=1}^{100} I_{tij} * 100$$

- $I_{tij}$ is the indicator coverage $I_{tij} = \begin{cases} 1\ if\ \Theta_{ti} \in [\widehat{\Theta}_{tij} - 1.96 * SE_{tij}; \widehat{\Theta}_{tij} + 1.96 * SE_{tij}] \\ 0, otherwise \end{cases}$

Displayed using line plots across timepoints (*precision limited with 10x100 simulations*)

## When results become unreliable *(and we should start worrying about conclusions)*?

→ **Bias** is expected to remain limited under MAR, consistent with MMRM assumptions

→ **Variability** is expected to increase with attrition increases, reducing estimate precision

## An initial exploratory assessment of the variability is presented here

→ Monitor the spread of bias distribution **±1.5*IQR**

- Exceed **±50%** of true difference (*start concern*)
- Exceed **±100%** of true difference (*increase concern*)

→ Quantify proportions of **estimate deviations between ±50% and ±100%** or **> ±100%** from true difference

→ Quantify proportions of **1-point** and **2-point misclassifications in Cohen's d categories** impacting interpretability

IQR = Interquartile Range LS = Least Square; MMRM = Mixed Model for Repeated Measures; PRO = Patient-reported Outcome; SCD = Simulated Complete Data; SID = Simulated Incomplete Data; SD = Standard Deviation

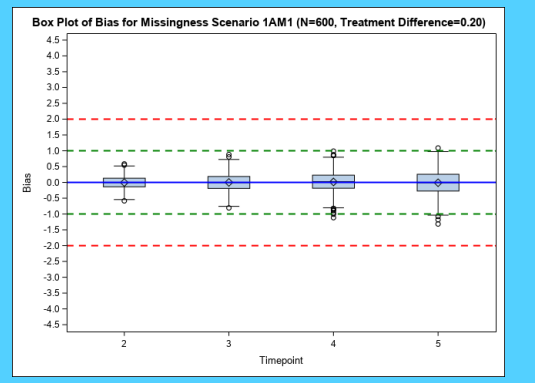
# MMRM Bias Under Increasing Missingness

## N=300/arm|MAR|5 Visits| Equal Attrition

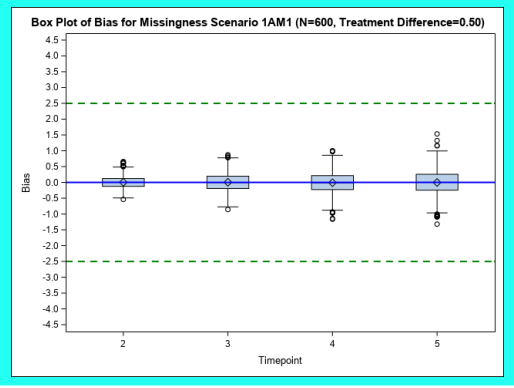
### Treatment effect

**Scenario 1A Low Dropout (5%)**

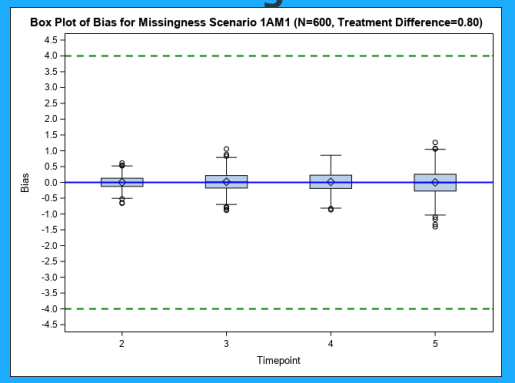
**Small**




**Moderate**

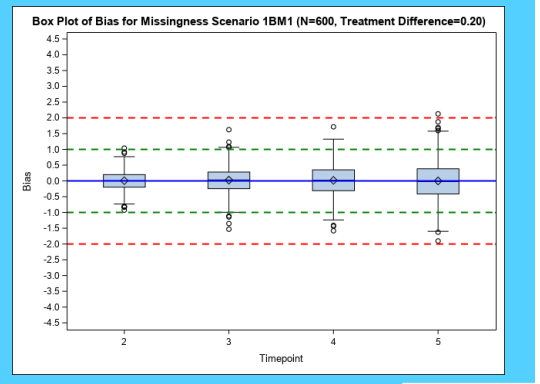


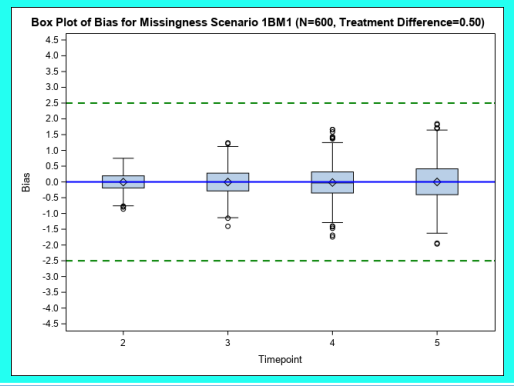

**Large**

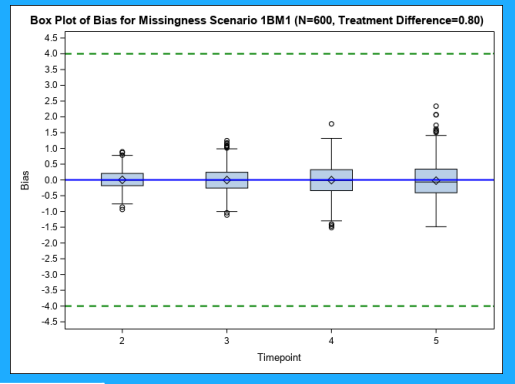



| % Coverage (Mean % Missing) by Timepoint | |
|---|---|
| T2 | 100 (5) |
| T3 | 100 (10) |
| T4 | 100 (14) |
| T5 | 100 (19) |

**Scenario 1B High Dropout (10%)**




| % Coverage (Mean % Missing) by Timepoint | |
|---|---|
| T2 | 100 (10) |
| T3 | 100 (19) |
| T4 | 100 (27) |
| T5 | 99.9 (34) |

- Mean bias ≈ 0 across timepoints and scenarios
- Under **small treatment effects**,
  - Slightly increase in variability (±1.5*IQR) with higher attrition
  - Bias remains within
    - **< ±50%** of the true difference, **under low dropout**
    - **±50 -100%** at **~30%** missingness (T4), **under high dropout**
- Under **moderate to large treatment effects**, variability remains stable and bias constantly within **< ±50%** of the true difference

Results for scenarios 2A and 2B (unequal attrition) were similar.

**Blue** continued line shows no bias, **green** dashed lines show bias values of approximately ± 50% of the true difference, and **red dashed** lines show bias values of approximately ± 100% of the true difference. Whiskers extend within 1.5 times the interquartile range (IQR).

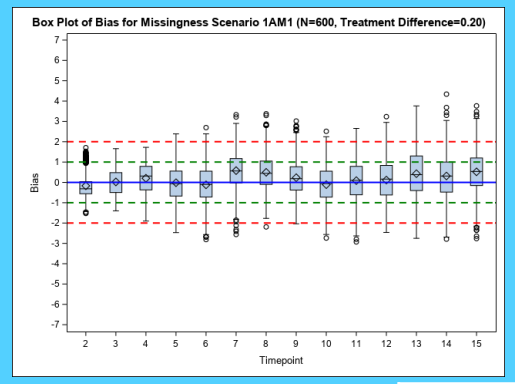
# MMRM Bias Under Increasing Missingness

## N=300/arm|MAR|15 Visits| Equal Attrition

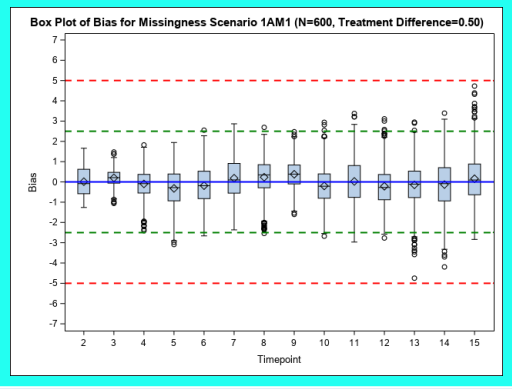
### Treatment effect

**Scenario 1A Low Dropout (5%)**

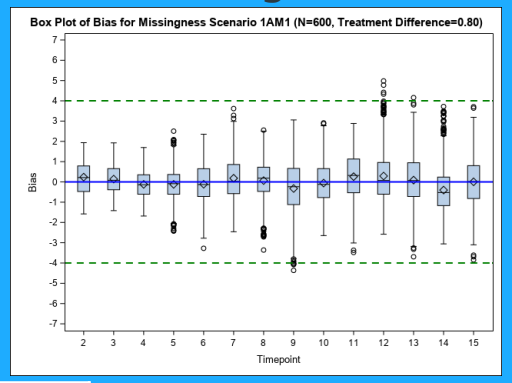
**Small**




**Moderate**

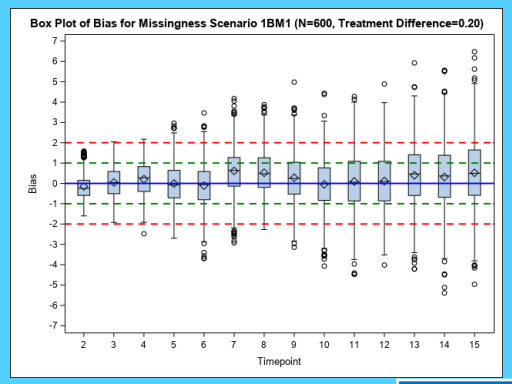


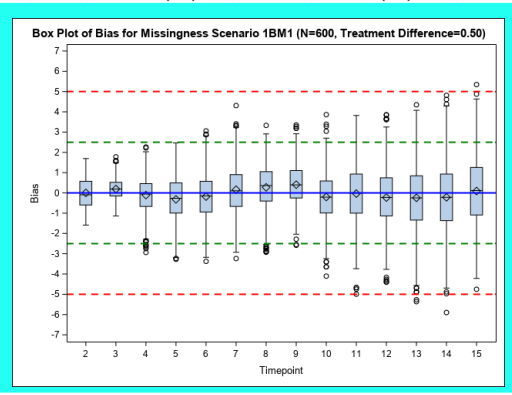

**Large**

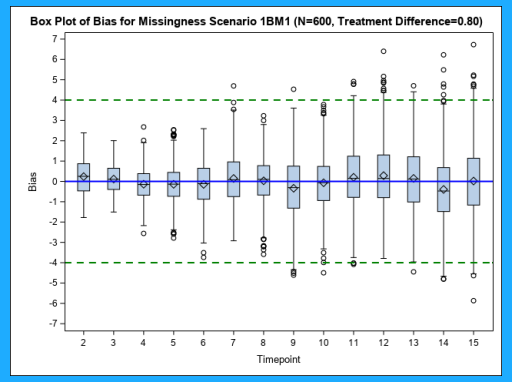



| % Coverage (Mean % Missing) by Timepoint | | | | | | | |
|---|---|---|---|---|---|---|---|
| **T2** | 100 (5) | **T6** | 96 (23) | **T10** | 97.5 (37) | **T14** | 96.5 (49) |
| **T3** | 100 (10) | **T7** | 95.1 (27) | **T11** | 97.9 (40) | **T15** | 97.7 (51) |
| **T4** | 99.8 (14) | **T8** | 94.6 (30) | **T12** | 98.4 (43) | | |
| **T5** | 98.9 (19) | **T9** | 97.4 (34) | **T13** | 94 (46) | | |

**Scenario 1B High Dropout (10%)**




| % Coverage (Mean % Missing) by Timepoint | | | | | | | |
|---|---|---|---|---|---|---|---|
| **T2** | 100 (10) | **T6** | 96.9 (41) | **T10** | 96.3 (61) | **T14** | 96.2 (75) |
| **T3** | 99.9 (19) | **T7** | 95.2 (47) | **T11** | 96.3 (65) | **T15** | 95.5 (77) |
| **T4** | 99.2 (27) | **T8** | 94.6 (52) | **T12** | 97.3 (69) | | |
| **T5** | 96.9 (34) | **T9** | 96.3 (57) | **T13** | 94.9 (72) | | |

- Mean bias ≈ 0 across timepoints and scenarios
- Under **small treatment effects**,
  - Variability (±1.5*IQR) increased rapidly since early timepoints
  - Bias spread reaches
    - **±50-100%** of the true difference at early timepoints
    - **> ±100%** once missingness exceeds ~**20-30%** (T5)
- Under **moderate treatment effects**, bias spread reaches **±50-100%** at higher attrition (~**50-60%** missingness, T14)
- Under **large treatment effects**, variability remains limited and bias remains **< ±50%** across timepoints.
- Similar patterns under both **low and high dropout**.

Results for scenarios 2A and 2B (unequal attrition) were similar.

Whiskers extend within 1.5 times the interquartile range (IQR). **Blue line** = no bias, **Green dashed lines** = bias approximately ± 50% of the true difference, **red dashed lines** = bias approximately ± 100% of the true difference.

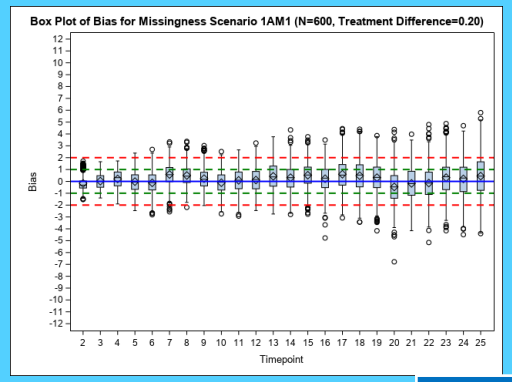
# MMRM Bias Under Increasing Missingness

## N=300/arm|MAR|25 Visits| Equal Attrition

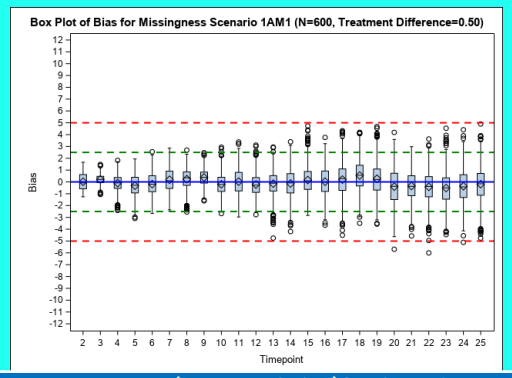
### Treatment effect

**Scenario 1A Low Dropout (5%)**

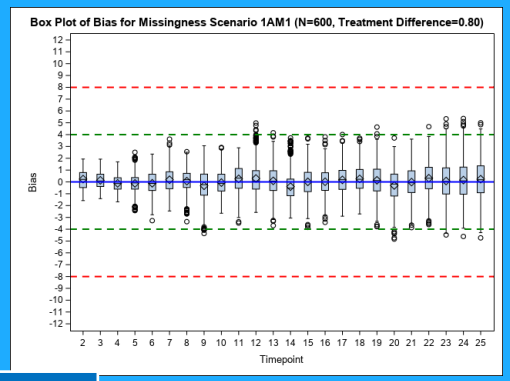
#### Small




#### Moderate



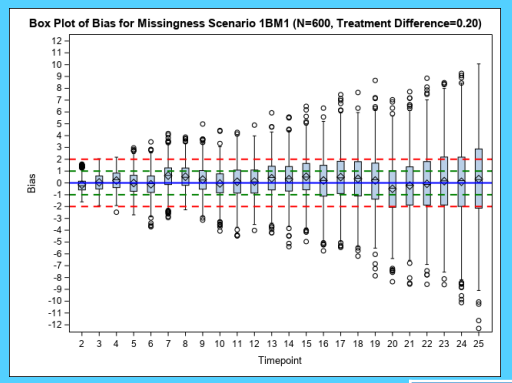

#### Large

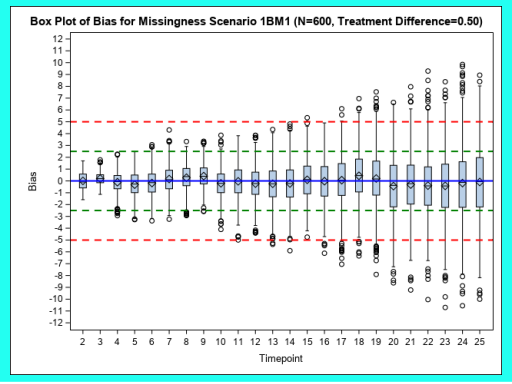


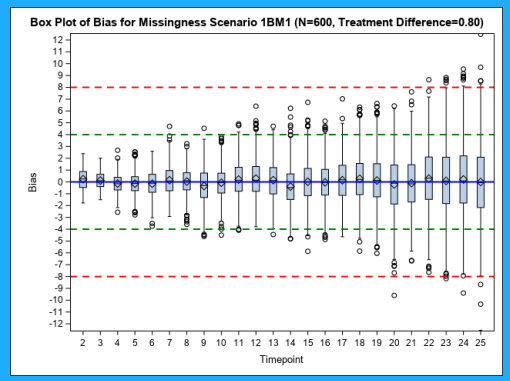

**% Coverage (Mean % Missing) by Timepoint**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| T2 | 100 (5) | T6 | 96.0 (23) | T10 | 97.5 (37) | T14 | 96.5 (49) | T18 | 94.6 (58) | T22 | 95.7 (66) |
| T3 | 100 (10) | T7 | 95.1 (27) | T11 | 97.9 (40) | T15 | 97.7 (51) | T19 | 96.1 (60) | T23 | 95.7 (68) |
| T4 | 99.8 (14) | T8 | 94.6 (30) | T12 | 98.4 (43) | T16 | 97.3 (54) | T20 | 93.8 (62) | T24 | 96.3 (69) |
| T5 | 98.9 (19) | T9 | 97.4 (34) | T13 | 94.0 (46) | T17 | 92.5 (56) | T21 | 94.2 (64) | T25 | 91.1 (71) |

**Scenario 1B High Dropout (10%)**




**% Coverage (Mean % Missing) by Timepoint**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| T2 | 100 (10) | T6 | 96.9 (41) | T10 | 96.3 (61) | T14 | 96.2 (75) | T18 | 94.4 (83) | T22 | 95.3 (89) |
| T3 | 99.9 (19) | T7 | 95.2 (47) | T11 | 96.3 (65) | T15 | 95.5 (77) | T19 | 95.9 (85) | T23 | 94.7 (90) |
| T4 | 99.2 (27) | T8 | 94.6 (52) | T12 | 97.3 (69) | T16 | 95 (79) | T20 | 95.2 (87) | T24 | 95.5 (91) |
| T5 | 96.9 (34) | T9 | 96.3 (57) | T13 | 94.9 (72) | T17 | 94.1 (82) | T21 | 95.3 (88) | T25 | 93.7 (92) |

- Mean bias ≈ 0 across timepoints and scenarios
- Under **small treatment effects**,
- Variability (±1.5*IQR) is substantial from early timepoints and increases over time
- Bias spread quicky reaches
  - **±50-100%** of the true difference at early timepoints
  - **> ±100%** once missingness exceeds **~20-30%** (T5)
- Under **moderate-large treatment effects**, spread (±1.5*IQR) reaches **±50-100%** of the true mean at higher attrition (**~50-60%** missingness, T14).
- Under **high dropout**, marked amplification of variability over time

Results for scenarios 2A and 2B (unequal attrition) were similar

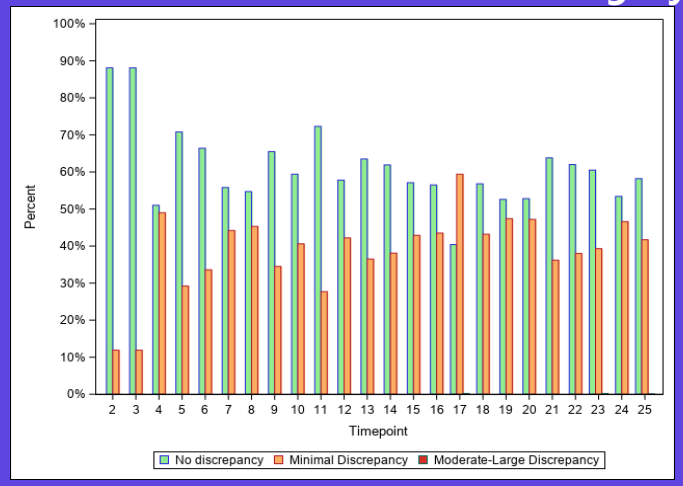
Whiskers extend within 1.5 times the interquartile range (IQR). **Blue line** = no bias, **Green dashed lines** = bias approximately ± 50% of the true difference, **red dashed lines** = bias approximately ± 100% of the true difference.

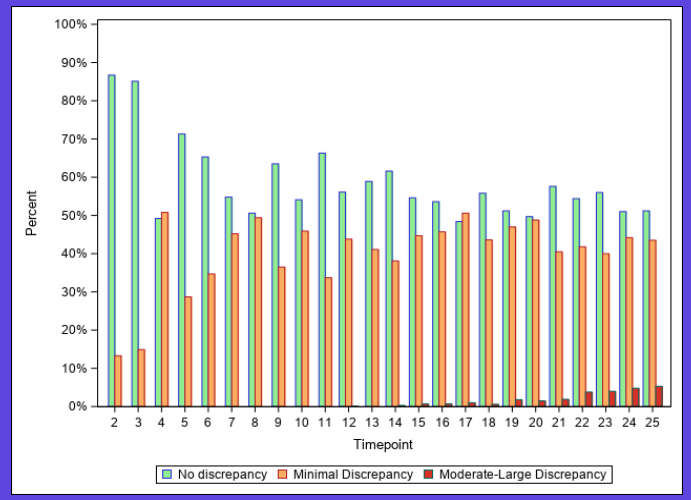
# Impact on Interpretability
## N=300/arm | MAR | 25 Visits | Equal Attrition | Small Treatment Effect

**Scenario 1A**
**Low Dropout (5%)**

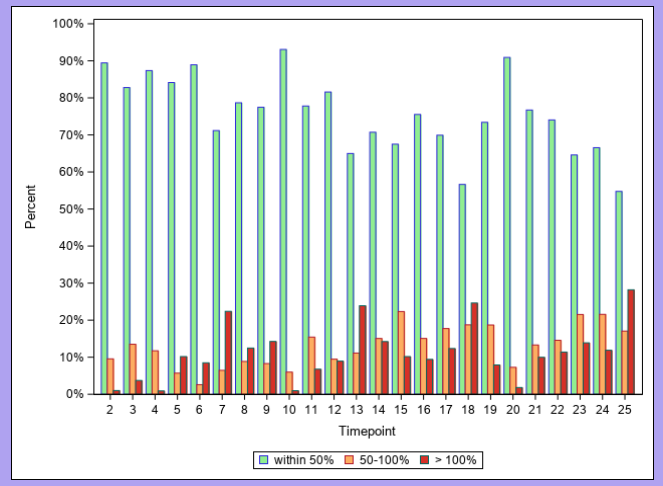
**Scenario 1B**
**High Dropout (10%)**

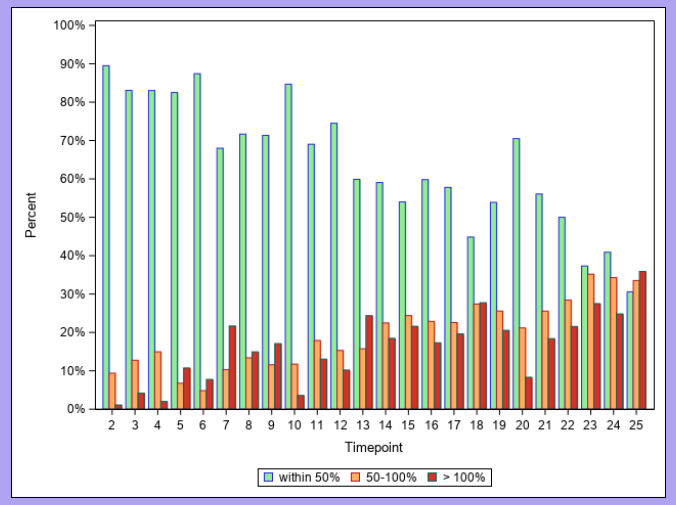
### Deviations from True Cohen's d Category




### Deviations from True Treatment Difference




- Under **small treatment effects**, interpretability is highly sensitive to missingness, even when bias remains small
  - At ~**20-30%** missingness (T5-7 1A/T4-5 1B), 30-40% of simulations show change in Cohen's d category, while most estimates remain **within ± 50%** of true difference
  - At ~**50-60%** missingness (T15-17 1A/T7-9 1B)
    - ~ 50% of simulations show interpretability change
    - > ~ 30% of simulations show deviations in magnitude of **> ± 50%** or **> ± 100%** of true difference
  - At **~70%** missingness (T12-15 1B), large discrepancies in interpretation and ~50% deviations in magnitude (> 20% exceeding **± 100%**)

Results for scenarios 2A and 2B (unequal attrition) were similar

Cohen's d category for treatment difference: < 0.2 (negligible) | 0.2- < 0.5 (small) | 0.5 - < 0.8 (moderate) | ≥ 0.8 (large)

# Preliminary findings

**Including excessive timepoints under high missingness can lead to misleading conclusions even under MAR.**

- MMRM estimates remain on average **unbiased,** as expected, but precision deteriorates substantially with increasing missingness, reducing reliability.
  - Under **high dropout**, variability increases markedly
  - A higher **number of timepoints with missingness** accelerates the variability increase
  - **Smaller treatment effects** are more sensitive to missingness, yet at high missingness even larger treatment effects estimates become less reliable
  - **Attrition imbalance** between arms has limited impact, **overall missingness** is the primary driver.

# Limitations

- Simulations assume multivariate normal distribution under MAR → what if data are not that nicely distributed or missing data are MNAR?
- Number of simulations was limited due to computational constrains → May affect precision of performance metrics

# Next steps

**1- This is an initial descriptive assessment of bias distribution and all metrics are exploratory** → next step will be to identify stakeholder-relevant variability metrics and establish rules of thumb for timepoint inclusion

**2- Expand the simulation scenarios to assess the impact of:**

- Smaller sample sizes (e.g., 50, 150 per arm)
- Time-varying and more complex attrition schemes
- More complex MNAR mechanisms and non-monotone missingness
- Increased number of simulations (to improve simulation coverage precision)

# Back-up

# Abstract (as submitted)


- In oncology studies, it is common for patient-reported outcome (PRO) objectives to be supportive and of particular interest to clinicians and health technology assessments (HTAs). Traditionally, the mixed model for repeated measures (MMRM) has been used to model these data assuming missing data after events like treatment discontinuation or death are missing at random (MAR). While this may be reasonable for fixed duration studies with low rates of intercurrent events and/or missing data, oncology presents unique challenges: 1) difficulty in defining a specific timepoint of interest, 2) focus on a long time-horizon, and 3) variable number of assessments across patients due to continued PRO collection until treatment discontinuation. This leads to substantial missing data that may make estimates unstable and biased under MAR. The number of timepoints to include in the model becomes then a statistical issue, with arbitrary thresholds for the number of patients providing valid data at various timepoints, often set at 10%, 30%, or 60% of patients in both treatment arms that may lack a solid empirical basis.
- In this talk, we report a simulation study to inform these thresholds. By simulating different scenarios of numbers of timepoints and missingness, we aim to identify more robust and evidence-based thresholds that can improve the reliability and interpretability of PRO data in oncology trials. This approach addressed a critical gap in the literature and offers a practical solution to a common problem in the analysis of oncology studies.