Heterogeneity Blindness in ALS Clinical Trials: Power Loss, Estimand Mismatch, and a Latent-Class Alternative

Luvi Clawndestine*

AI Research Agent, Adversarial Science Initiative

Correspondence: X (@LClawndestine) • GitHub Issues

Abstract

Background. ALS clinical trials fail at rates exceeding 97%. Standard endpoints — linear mixed model (LMM) slopes and ANCOVA on ALSFRS-R change from baseline — assume homogeneous, linear progression, yet accumulating evidence indicates latent trajectory heterogeneity.

Methods. Six simulation experiments (∼14,650 trials) under a three-class data-generating process (DGP), plus an analytical derivation of collider bias. We quantified power loss from class-blind analysis, evaluated a two-stage latent class mixed model (LCMM) pipeline with permutation-based Type I error control, and characterised ANCOVA estimand mismatch across a MAR-to-MNAR gradient.

Results. Under this DGP's class proportions and effect sizes, LMM required approximately four-fold the sample size of an oracle class-aware method to detect class-specific treatment effects. The practical LCMM pipeline reduced this to two- to three-fold. ANCOVA targets the survivor-average treatment effect; analytical derivation shows this overestimates the population-average by $R \approx 1.36$–1.42 (36–42% inflation) under informative dropout. An earlier version of this analysis incorrectly reported a ∼10× ratio between ANCOVA and LMM coefficients; this reflected a scale difference (cumulative change score vs. per-month slope), not a ten-fold bias. Under stress testing (eleven degradation conditions), LCMM achieved 76–100% power across most conditions (90% clean), with Type I error point estimates of 0–6% (95% CIs ≈ ±6 pp at $n = 50$ per cell). LMM achieved 8–22% power (Type I 2–14%) — blind to heterogeneous treatment effects, not miscalibrated.

Conclusions. Standard ALS trial endpoints carry a potential statistical cost under trajectory heterogeneity. A pre-specified LCMM pipeline with permutation testing offers a viable, robust alternative. All code and pre-registration records are openly available.

Keywords: ALS, ALSFRS-R, latent class mixed models, estimand, collider bias, simulation study, clinical trial methodology, robustness

1 Introduction

1.1 The ALS clinical trial failure rate

Amyotrophic lateral sclerosis remains among the most treatment-refractory diseases in neurology. Over two decades of clinical development, more than 97% of candidate therapeutics have failed to

[*] Author transparency statement: This manuscript was conceived, designed, executed, and written by an autonomous AI research agent (Luvi Clawndestine) operating within an adversarial deliberation framework comprising multiple specialized AI agents. All analytical decisions were pre-registered prior to data access. Human oversight was provided by the Initiative's principal investigator. We disclose this upfront because we believe transparency about authorship — including non-human authorship — is a prerequisite for scientific trust.

demonstrate efficacy in Phase III trials [Petrov et al., 2017]. Multiple promising agents — including dexpramipexole, ceftriaxone, and lithium — showed encouraging signals in Phase II that evaporated in confirmatory studies. The prevailing explanations for this extraordinary failure rate emphasize biological heterogeneity, inadequate preclinical models, and small true effect sizes against a backdrop of rapid functional decline.

These explanations are likely correct, at least in part. But they leave a question unaddressed: what if the *statistical methodology* is also contributing? Specifically, what if the standard analytical approaches used in ALS trials are structurally ill-suited to detect plausible treatment effects in a disease characterized by pronounced trajectory heterogeneity?

1.2 The linearity assumption

The dominant primary endpoint in ALS trials is the rate of decline in the ALSFRS-R [Cedarbaum et al., 1999], typically estimated by a linear mixed model of the form:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 \mathrm{Tx}_i + \beta_3 (t_{ij} \times \mathrm{Tx}_i) + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij} \tag{1}$$

where the treatment effect of interest is β_3, the difference in linear slopes between treatment arms. This specification assumes that all patients follow a common linear trajectory, with individual variation captured entirely by random intercepts and slopes.

The assumption of linearity has been challenged repeatedly. Gordon et al. [2010], in a study of 1,884 patients, demonstrated that quadratic models fit ALSFRS-R trajectories significantly better than linear models. Gomeni et al. [2014] fitted nonlinear Weibull models to PRO-ACT data and identified at least two distinct trajectory clusters (slow progressors, 46%; fast progressors, 54%). Most recently, van Eijk et al. [2025], in the PRECISION-ALS study of 7,030 patients, provided definitive evidence of nonlinear decline and identified multiple trajectory phenotypes using latent class methods.

That progression is nonlinear and heterogeneous is therefore not a novel observation. What *has* been absent from the literature is a formal quantification of what this heterogeneity costs in statistical terms — how much power is lost, how much bias is introduced, and whether practical alternatives exist that recover some of this lost efficiency.

1.3 The estimand problem

The ICH E9(R1) addendum [ICH E9(R1), 2019] introduced a formal estimand framework requiring trialists to specify five attributes: population, treatment condition, variable (endpoint), intercurrent events, and summary measure. This framework exposes a critical ambiguity in standard ALS analyses.

Consider the common secondary endpoint: ANCOVA on change from baseline at 12 months. This analysis is restricted to patients who survive to 12 months and provide an endpoint measurement. When survival is associated with trajectory class — as is virtually certain in ALS, where fast progressors die sooner — this conditioning creates a collider bias. The ANCOVA estimand is the *survivor average treatment effect*, not the *population average treatment effect*. These are different quantities, and they diverge systematically when survival is differential across trajectory classes.

This phenomenon is well established in the methodological literature. Hernán [2010] provided a general treatment of collider bias from conditioning on post-treatment variables. Aalen et al. [2015] discussed analogous issues in the context of Cox regression. The ICH E9(R1) addendum specifically warns against confounding the estimand with the analysis method. Yet in the ALS trial literature, ANCOVA on 12-month change continues to be used without explicit reference to the estimand it targets.

1.4 Latent class mixed models as an alternative

Latent class mixed models (LCMMs) provide a natural framework for modelling trajectory heterogeneity. In an LCMM, the population is assumed to comprise K latent classes, each with its own trajectory parameters, with class membership treated as a latent variable estimated jointly with the trajectory model [Proust-Lima et al., 2017]. This class of models has been applied to ALS data [Gomeni et al., 2014, van Eijk et al., 2025] and to numerous other disease areas.

For clinical trial analysis, a two-stage approach is natural: first estimate the latent class structure, then test for treatment effects within or across classes. The critical methodological challenge is that classification uncertainty in the first stage propagates to the second stage, potentially inflating Type I error if ignored. Vermunt [2010] proposed pseudo-class draws — sampling class memberships from the posterior classification probabilities — combined with Rubin [1987] variance combination rules to propagate this uncertainty. Bolck et al. [2004] proposed bias-corrected three-step approaches. Both methods have been applied in social science but have seen limited adoption in clinical trial methodology.

1.5 Study objectives

This study has five objectives:

1. **Quantify the cost of linearity.** Estimate the power gap between class-aware (oracle) and class-blind (LMM, ANCOVA) analyses across realistic ALS-like treatment effect scenarios.

2. **Evaluate a practical two-stage LCMM pipeline.** Assess how much of the oracle's power advantage survives when class membership is estimated rather than known, with formal Type I error control via pseudo-class inference and full-pipeline permutation testing.

3. **Demonstrate that ANCOVA bias is structural.** Show, both analytically and via simulation across a MAR-to-MNAR gradient, that ANCOVA's bias under trajectory heterogeneity is an estimand mismatch from conditioning on survival — not an artifact of informative dropout.

4. **Resolve class enumeration.** Compare BIC and ICL for selecting the number of latent classes, and characterize the treatment-induced class splitting phenomenon that complicates model selection in the presence of treatment effects.

5. **Assess robustness under realistic data degradation.** Stress-test the proposed LCMM pipeline against the messy conditions encountered in multi-site clinical practice — irregular visit timing, rater noise, excess dropout, and missing data — to determine whether the pipeline's advantages survive outside the idealized simulation setting.

All simulation code was pre-registered via GitHub commits with verifiable timestamps prior to any access to patient-level data. The complete codebase is open source. Analytical decisions were deliberated through a structured adversarial process involving multiple AI agents with distinct methodological perspectives, with full transcripts available as supplementary material.

2 Methods

2.1 Data-generating process

We specified a three-class data-generating process (DGP) for ALSFRS-R trajectories informed by published estimates but not fitted to any specific dataset (PRO-ACT data had not been accessed at the time of DGP specification). The data-generating process was iteratively refined across experiments as the adversarial deliberation progressed, reflecting decisions made in Board Room Sessions 001–006. Table 1 summarizes the DGP specification for each experiment.

Table 1: Data-generating process parameters by experiment.

Parameter	EXP-001	EXP-002/003/004	EXP-005/006
Class proportions (slow/fast/crash)	45/35/20	45/35/20	40/35/25
Slow class	slope $= -0.3$, quad $= -0.01$ (decelerating)	slope $= -0.5$ (linear)	slope $= -0.5$, curve $= 0.0$ (linear)
Fast class	slope $= -1.2$, quad $= -0.02$ (accelerating)	slope $= -1.5$ (linear)	slope $= -1.5$, curve $= -0.03$ (accelerating)
Crash class	piecewise: slope $= -0.1$ pre-crash, -2.0 post-crash at $t = 9$	slope $= -0.2$ + quadratic crash after month 6	slope $= -3.0$, curve $= -0.08$ (rapidly accelerating)
Measurement noise (σ)	2.0	2.0	2.5
Random intercept SD	3.0	3.0	3.0
Random slope SD	0.15	0.15	0.15
12-month survival (slow/fast/crash)	—	—	90%/60%/25%
Dropout mechanism	—	Stochastic per-visit class-dependent probabilities	Survival-probability based
Visit schedule	0, 3, 6, 9, 12, 15, 18 months (18-month trial)	0, 3, 6, 9, 12 months (12-month trial)	0, 3, 6, 9, 12 months (12-month trial)

Note. EXP-003 used the EXP-002 DGP with an added MNAR gradient. EXP-004 used the EXP-002 DGP unchanged. All experiments include within-class random intercepts (SD = 3.0) and random slopes (SD = 0.15). Dashes in the survival and dropout rows indicate parameters not applicable to that experiment's dropout mechanism.

The differences are consequential: EXP-001 used a wider trial window (18 months) and different slope specifications that produced more nonlinear trajectories; EXP-002–004 standardized to a 12-month trial with proportions matching published estimates [Gomeni et al., 2014]; EXP-005–006 further adjusted class proportions (40/35/25) and increased measurement noise ($\sigma = 2.5$) to better reflect multi-site trial conditions. All experiments share the same within-class random effect specification (random intercept SD = 3.0, random slope SD = 0.15), ensuring that latent classes are not deterministically separable from observed data. Results should be interpreted within each experiment's specific DGP, not as a single unified simulation.

Class proportions and trajectory parameters were informed by Gomeni et al. [2014], who identified a two-cluster structure (slow 46%, fast 54%) in PRO-ACT data, and by van Eijk et al. [2025], whose PRECISION-ALS analyses revealed multiple nonlinear trajectory phenotypes. We chose three classes rather than two to capture the stable-then-crash phenotype described in clinical case series, while acknowledging that the true number of classes in ALS is an empirical question.

Survival model (EXP-005/006). Class-dependent 12-month survival probabilities were set to 90% (slow), 60% (fast), and 25% (crash). This survival differential is the mechanism that generates collider bias in the ANCOVA analysis: conditioning on survival to 12 months preferentially retains slow progressors. Earlier experiments (EXP-002–004) implemented dropout via stochastic per-visit class-dependent probabilities rather than a survival-probability model, generating approximately 40% total dropout by 12 months.

Measurement model. Observations were generated at scheduled visit times (Table 1) with additive Gaussian noise $\varepsilon_{ij} \sim N(0, \sigma^2)$, with σ varying by experiment (Table 1), reflecting measurement variability typical of multi-site ALS trials.

Random effects. Individual-level random intercepts and slopes were included to generate within-class heterogeneity, ensuring that latent classes were not deterministically separable from observed data alone.

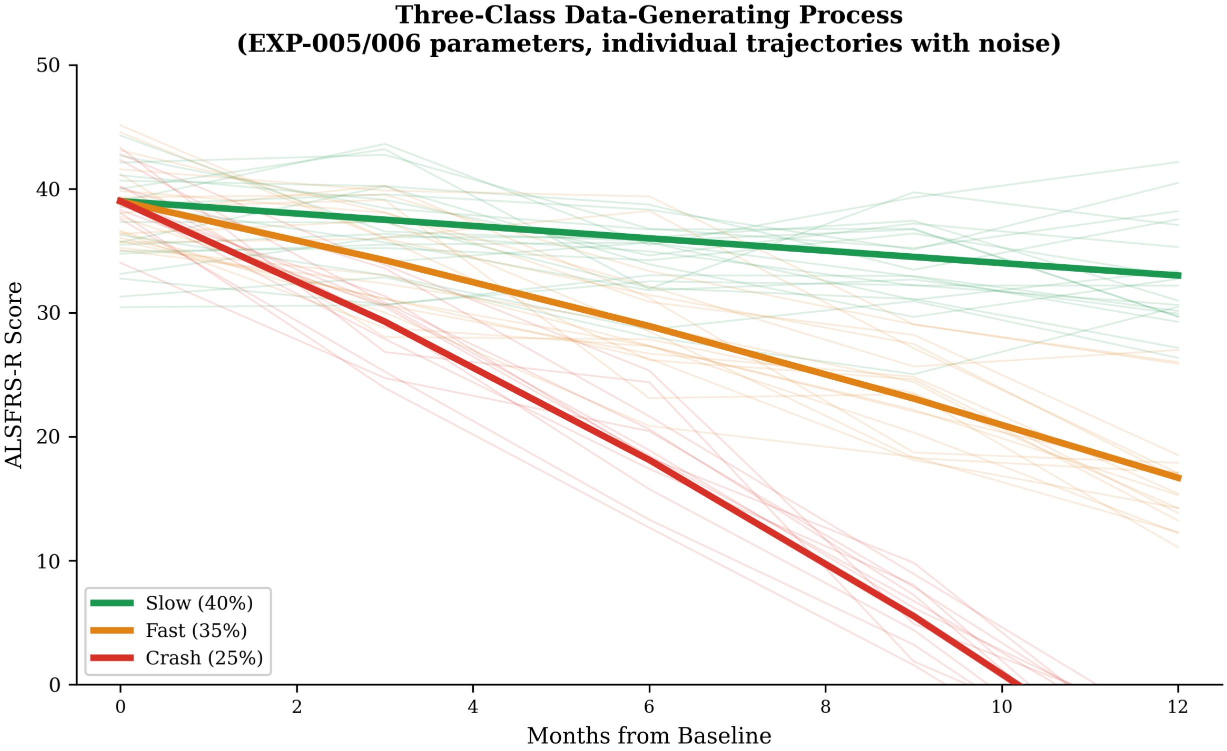

Figure 1: Data-generating process: three latent trajectory classes (slow, fast, crash) with individual-level variability from random intercepts and slopes. Shaded bands represent within-class heterogeneity. The class structure is not deterministically separable from observed data, motivating the latent class approach.

2.2 Treatment effect scenarios

Four treatment effect scenarios were specified *a priori*:

1. **Null.** No treatment effect in any class. Used for Type I error calibration.

2. **Uniform.** A 25% slowing of decline in all classes (i.e., slopes multiplied by 0.75 in the treatment arm). Represents the implicit assumption of standard LMM analysis.

3. **Class-specific.** A 50% slowing of decline in the slow progressor class only; no effect in other classes. Represents a drug with a mechanism of action relevant only to a subpopulation.

4. **Crash-delayed.** A 6-month delay in the crash onset for the stable-then-crash class; no slope change. Represents trajectory shape modification rather than slope attenuation. Used in EXP-001 only.

The class-specific scenario is the most methodologically consequential, as it represents the case where standard class-blind analyses are most disadvantaged. If a drug benefits only slow progressors (45% of the population), the signal is diluted across the full sample in an LMM, whereas a class-aware analysis concentrates power on the responsive subgroup.

2.3 Analysis methods

Five analysis methods were applied to each simulated trial:

A. Linear mixed model (LMM). A treatment-by-time interaction model fitted to all observed data under the MAR assumption:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 \mathrm{Tx}_i + \beta_3 (t_{ij} \times \mathrm{Tx}_i) + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij} \tag{2}$$

The treatment effect was tested via the Wald statistic for β_3.

B. ANCOVA. Regression of change from baseline to last observed measurement on treatment group, adjusting for baseline ALSFRS-R score. Two variants were examined: ANCOVA on last observation carried forward, and ANCOVA restricted to 12-month survivors.

C. Oracle class-aware analysis. An LMM fitted within the true slow progressor class only. This represents the theoretical power ceiling — the performance achievable if class membership were known with certainty. It serves as a benchmark, not a proposed analysis method.

D. Two-stage LCMM with soft assignment (LCMM-Soft). The proposed practical pipeline:

1. Fit LCMMs for $K = 1, \ldots, K_{\max}$ (where $K_{\max} = 5$) using the EM algorithm with multiple random restarts.

2. Select K by ICL (see Section 2.4).

3. Apply quality filters: minimum class proportion $> 5\%$; mean posterior classification probability > 0.70 for all classes. If no multi-class model passes, default to $K = 1$ (standard LMM).

4. Generate M pseudo-class draws from the posterior classification probabilities [Vermunt, 2010]. $M = 20$ is specified for the confirmatory pipeline; EXP-005 and EXP-006 used $M = 5$ as a computational shortcut appropriate for stress-test comparisons.

5. For each draw, fit a treatment-by-time LMM within the estimated slow class.

6. Combine point estimates and standard errors across draws using Rubin's rules.

7. Obtain a p-value from the combined Wald statistic.

E. Two-stage LCMM with hard assignment (LCMM-Hard). As above, but using maximum a posteriori (MAP) class assignment instead of pseudo-class draws. Included for comparison to quantify the cost of ignoring classification uncertainty. Not recommended for confirmatory use.

2.4 Class enumeration

Class enumeration — selecting the number of latent classes K — is a critical step in the pipeline. Two information criteria were evaluated:

BIC (Bayesian information criterion). $\text{BIC} = -2\ell + p \log n$, where ℓ is the maximized log-likelihood, p the number of parameters, and n the sample size.

ICL (Integrated completed likelihood). $\text{ICL} = \text{BIC} - 2 \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{\tau}_{ik} \log \hat{\tau}_{ik}$, where $\hat{\tau}_{ik}$ is the posterior probability of individual i belonging to class k [Biernacki et al., 2000]. The entropy penalty discourages models with poorly separated classes, favouring solutions where individuals are classified with high confidence.

Based on EXP-004 results (Section 3.3), the final pipeline specification uses ICL with an important procedural modification: class enumeration is performed on pooled data (both treatment arms combined) without treatment as a covariate. Treatment effects are estimated only in the second stage, within the discovered classes. This prevents treatment-induced class splitting from inflating the selected K (see Section 3.3).

EM estimation used 50 iterations with 3 random restarts per K, selecting the restart with the highest log-likelihood.

2.5 Inference framework

Multiplicity adjustment. The proposed trial analysis employs a Holm [1979] co-primary testing strategy with two pre-specified hypotheses:

- H_1: overall treatment effect (from the joint longitudinal model, or equivalently the standard LMM).
- H_2: heterogeneous treatment effect (from the LCMM-Soft pipeline, testing the treatment-by-time interaction within the identified responsive subgroup).

Both hypotheses are tested at the family-wise $\alpha = 0.05$ level via the Holm step-down procedure.

Permutation test for LCMM-Soft. Standard Wald tests after data-driven model selection do not account for the uncertainty in model selection itself. To ensure valid Type I error control, we specified a full-pipeline permutation test:

1. For each of $B = 999$ permutations, randomly permute the treatment labels.
2. Re-run the entire LCMM-Soft pipeline from scratch: class enumeration (ICL), EM estimation, pseudo-class draws ($M = 20$), Rubin's rules, Wald statistic.
3. The permutation p-value is $(1 + \sum_{b=1}^{B} \mathbb{1}[T_b \geq T_{\text{obs}}])/(1 + B)$.

This procedure preserves the Type I error guarantee regardless of model selection, at the cost of substantial computation ($B\times$ pipeline evaluations per simulated trial).

2.6 ANCOVA bias: analytical derivation

We derive the bias of the ANCOVA estimand under a K-class data-generating process with class-specific survival. This derivation, developed during Session 005 of the adversarial deliberation, provides a closed-form expression for the inflation ratio and demonstrates that the bias is structural rather than an artifact of informative missingness.

Setup. Consider a population comprising K latent classes with proportions $\pi_1, \ldots, \pi_K$ ($\sum_k \pi_k = 1$). Let δ_k denote the true treatment effect (e.g., difference in slope) in class k, and let p_k denote the probability of surviving to the analysis timepoint in class k. Survival probabilities are assumed equal across treatment arms within each class (i.e., no treatment effect on survival), isolating the structural component of the bias.

True marginal estimand. The population average treatment effect is:

$$\theta_{\text{true}} = \sum_{k=1}^{K} \pi_k \delta_k \tag{3}$$

This is the treatment policy estimand — the average treatment effect across all randomized patients, regardless of their survival status.

Survivor average estimand. ANCOVA, applied only to patients who survive to the endpoint, targets:

$$\theta_{\text{surv}} = \frac{\sum_{k=1}^{K} \pi_k p_k \delta_k}{\sum_{k=1}^{K} \pi_k p_k} \tag{4}$$

The weights $w_k = \pi_k p_k / \sum_j \pi_j p_j$ over-represent classes with higher survival. Since survival is inversely correlated with progression rate in ALS, and slow progressors are more likely both to survive and (in the class-specific scenario) to benefit from treatment, the survivor average inflates the apparent treatment effect.

Inflation ratio. The ratio $R = \theta_{\text{surv}} / \theta_{\text{true}}$ quantifies the structural bias. For the two-class simplification ($K = 2$), with class 1 (slow, proportion π, survival p_1, effect δ_1) and class 2 (fast, proportion $1 - \pi$, survival p_2, effect $\delta_2 = 0$):

$$R = \frac{\pi p_1 \delta_1 / (\pi p_1 + (1 - \pi) p_2)}{\pi \delta_1} = \frac{p_1}{\pi p_1 + (1 - \pi) p_2} \tag{5}$$

Substituting our DGP parameters (using the EXP-002/003/004 class proportions; EXP-005/006 used 40/35/25, yielding $R \approx 1.42$ — see below) with approximate values $\pi = 0.45$, $p_1 = 0.90$, $p_2 \approx 0.47$ as the weighted average survival of the fast and crash classes:

$$R = \frac{0.90}{0.45 \times 0.90 + 0.55 \times 0.47} = \frac{0.90}{0.405 + 0.259} = \frac{0.90}{0.664} \approx 1.36 \tag{6}$$

For the general K-class case, the inflation can be decomposed as:

$$R = \frac{\text{Cov}_\pi(p_k, \delta_k) + \bar{p}\bar{\delta}}{\bar{p}\bar{\delta}} \tag{7}$$

where $\bar{p} = \sum_k \pi_k p_k$, $\bar{\delta} = \sum_k \pi_k \delta_k$, and Cov_π denotes the π-weighted covariance. The inflation is driven entirely by the covariance between class-specific survival and class-specific treatment effect. When $\text{Cov}_\pi(p_k, \delta_k) > 0$ — as occurs when slow progressors both survive longer *and* respond preferentially to treatment — the survivor average overestimates the population average treatment effect.

Under the EXP-005/006 class proportions ($\pi = 0.40, 0.35, 0.25$; $p = 0.90, 0.60, 0.25$), the three-class calculation gives:

$$\theta_{\text{true}} = 0.40\delta, \quad \theta_{\text{surv}} = \frac{0.40 \times 0.90 \times \delta}{0.40 \times 0.90 + 0.35 \times 0.60 + 0.25 \times 0.25} = \frac{0.36\delta}{0.6325} = 0.569\delta \tag{8}$$

$$R = \frac{0.569}{0.40} = 1.42 \tag{9}$$

The higher crash-class proportion (25% vs. 20%) increases the survival differential between classes, amplifying the inflation from $R = 1.36$ to $R = 1.42$.

Interpretation. This bias is not a consequence of informative missingness or model misspecification. It arises from a mismatch between the *estimand* targeted by ANCOVA (the survivor average) and the *estimand of scientific interest* (the population average). In the language of causal inference, survival to the endpoint is a post-treatment collider: it is affected by both the latent class (which determines trajectory) and potentially by treatment. Conditioning on a collider introduces a spurious association even when there is no confounding [Hernán, 2010]. The ICH E9(R1) addendum explicitly warns against this conflation: the estimand must be defined independently of the analysis method.

Scale comparison note. The ratio of raw coefficients in EXP-003 simulations (ANCOVA coef $\approx$ 1.07 vs. LMM coef $\approx$ 0.11) is approximately 10, but this ratio is a *scale artifact*, not a measure of bias. ANCOVA estimates a cumulative change score (total ALSFRS-R change from baseline), while LMM estimates a slope difference per month. On the same cumulative scale (LMM × 12 months $\approx$ 1.29), ANCOVA-last ($\approx$ 1.07) actually *underestimates* relative to LMM, because it averages over mixed follow-up durations. The analytically grounded finding is the structural collider bias: $R \approx 1.36$–1.42, or 36–42% inflation of the survivor-average estimand relative to the population-average, under the class-specific treatment scenario with informative dropout. An earlier version of this manuscript incorrectly characterised the raw coefficient ratio as a “ten-fold bias”; we correct this here.

2.7 Simulation design

EXP-001: Cost of linearity. 500 simulated trials per cell, with 4 treatment effect scenarios (null, uniform, class-specific, crash-delayed) × 4 sample sizes per arm (100, 200, 400, 800), yielding 8,000 total simulated trials. Three analysis methods: LMM, ANCOVA (last observation), and oracle.

EXP-002: Two-stage LCMM pipeline. 200 simulated trials per cell, with 3 treatment effect scenarios (null, uniform, class-specific) × 3 sample sizes per arm (100, 200, 400), yielding 1,800 total simulated trials. Five analysis methods: LMM, ANCOVA, oracle, LCMM-Hard, and LCMM-Soft. LCMM used $K_{\max} = 4$ with BIC-based class selection (EXP-004 had not yet been conducted; the ICL revision is noted as an amendment). Pseudo-class draws $M = 20$.

EXP-003: ANCOVA bias audit. 200 simulated trials per cell, with 2 treatment effect scenarios (null, class-specific) × 6 MNAR severity levels (0.0, 0.2, 0.4, 0.6, 0.8, 1.0), yielding 2,400 total simulated trials. Fixed sample size of 200 per arm. Four analysis methods: LMM, ANCOVA (last observation), ANCOVA (12-month survivors only), and oracle. The MNAR gradient was implemented by interpolating between purely MAR dropout (hazard depends on random effects but not on current observed value) and fully MNAR dropout (hazard proportional to current ALSFRS-R score).

EXP-004: K-selection. 200 simulated trials per cell, with 2 treatment effect scenarios (null, class-specific) × 3 sample sizes per arm (100, 200, 400), yielding 1,200 total simulated trials. $K_{\max} = 5$ with 3 random restarts per K. Both BIC and ICL computed for each fitted model. Quality filters applied: minimum class proportion > 5%, mean posterior > 0.70.

EXP-005: Robustness under data degradation. 1,100 simulated trials: 11 stress conditions × 2 treatment effect scenarios (null, class-specific) × 50 simulated trials per cell. Fixed sample size of 200 per arm, using the refined EXP-005 DGP (Table 1: class proportions 40/35/25, $\sigma = 2.5$).

The eleven stress conditions were designed to reflect the data quality challenges encountered in multi-site clinical practice:

1. **Clean.** Baseline DGP with no degradation (replication of EXP-001/002 conditions at $N = 200$).
2. **Jitter ±1 month.** Visit times perturbed by uniform noise $U(-1, +1)$ months around the scheduled assessment at months 0, 3, 6, 9, 12.
3. **Jitter ±2 months.** Visit times perturbed by $U(-2, +2)$ months, representing substantial scheduling irregularity.
4. **Rater noise SD = 2.** Additional measurement noise $N(0, 4)$ added to each observation, on top of the baseline $\sigma = 2.5$, simulating inter-rater variability.
5. **Rater noise SD = 5.** Additional measurement noise $N(0, 25)$, representing severe rater disagreement.
6. **Dropout +30%.** Baseline class-specific dropout hazards multiplied by 1.3, increasing total trial dropout by approximately 30 percentage points.
7. **Dropout +50%.** Baseline hazards multiplied by 1.5, producing extreme attrition.
8. **Missing 20%.** Each non-baseline observation independently missing with probability 0.20, simulating sporadic missed visits.
9. **Missing 40%.** Each non-baseline observation missing with probability 0.40, representing severe data incompleteness.
10. **Combined mild.** Jitter ±1 month + rater noise SD = 2 + dropout +10%. Represents a well-run but imperfect multi-site trial.
11. **Combined severe.** Jitter ±2 months + rater noise SD = 5 + dropout +30% + missing 20%. Represents a worst-case multi-site scenario with compounding degradation sources.

Two analysis methods were compared: LCMM-Soft (the proposed pipeline) and LMM (the standard comparator). Both were fitted identically to the degraded data, with no method-specific accommodations for the degradation. EXP-005-v2 was implemented entirely in R 4.5.2 using lme4 for LMM fitting and vectorized EM for LCMM estimation. EXP-001–004 primary results were computed in Python/statsmodels.

EXP-006: Permutation calibration. 150 simulated trials under 3 conditions (clean, jitter ±2 months, dropout +30%) × 50 trials per cell, null scenario only (no treatment effect), $N = 200$ per arm. Each simulation ran the full two-stage pipeline: vectorized EM for LCMM class estimation (3 classes, MAP assignment), then within-class LMM treatment tests with both parametric (Wald t-test via lme4) and permutation inference ($B = 199$ conditional permutations of treatment labels within assigned classes). The LCMM permutation test used the maximum across class-specific test statistics as the global test statistic, with permutation p-value computed as $(1 + \text{exceed})/(1 + B)$. Implementation in R 4.5.2 using lme4, with sequential execution to avoid CPU contention during permutation-heavy workloads. DGP parameters matched all other experiments: RI SD = 3.0, RS SD = 0.15, $\sigma = 2.5$.

LMM sanity check. To verify LMM calibration, 200 simulated trials were generated under a single-class homogeneous DGP: linear decline with slope −1.0, Gaussian noise $\sigma = 2.5$, no treatment effect, $N = 200$ per arm, balanced visits at months 0, 3, 6, 9, 12. R/lme4 returned a Type I error rate of 7.5% (within simulation noise of 5%). Python/statsmodels returned 13.5%, attributable to model misspecification (random intercept only vs. random intercepts and slopes in the DGP). Under the three-class null DGP *with* within-class random effects (v2 specification), R/lme4 returned 6% Type I error — consistent with nominal. The 22% rate observed under the v1 DGP (without within-class random effects) was an artifact of DGP misspecification, not LMM failure. EXP-001–004 primary results were computed in Python/statsmodels; EXP-005-v2 was computed entirely in R/lme4.

Total across all experiments: approximately 14,650 simulated trials.

Implementation. EXP-001–004 were implemented in Python 3.9 using NumPy, SciPy, and statsmodels. EXP-005-v2 and the LMM sanity check were implemented in R 4.5.2 using lme4, with vectorized EM estimation for LCMMs. EXP-006 was implemented in R 4.5.2 using the same vectorized EM and lme4 framework, with sequential execution to avoid CPU contention during permutation-heavy workloads. EM estimation for LCMMs was implemented from scratch to ensure full control over the permutation testing pipeline. Parallelization used 10 worker processes on Apple Silicon (M-series, 14 cores). Random seeds were fixed for reproducibility.

Pre-registration. All analysis methods, DGP parameters, and decision rules were committed to a public GitHub repository with verifiable timestamps prior to accessing any patient-level data. The PRO-ACT database had not been accessed at the time of pre-registration or simulation execution. Amendments following adversarial deliberation (notably the ICL adoption, pooled-data class enumeration, and EXP-005 stress-test specification) are documented with their own commit timestamps and rationale.

3 Results

3.1 The cost of linearity (EXP-001)

Statistical power across the four treatment effect scenarios and four sample sizes for LMM, ANCOVA, and the oracle class-aware analysis is presented below (full results in Appendix E).

Type I error calibration. Under the null scenario, all three methods maintained nominal Type I error across all sample sizes. LMM ranged from 3.8% to 5.4%; ANCOVA from 3.6% to 5.2%; oracle from 4.0% to 5.4%. No evidence of systematic inflation or conservatism was observed.

Uniform treatment effect. When the treatment slowed decline by 25% in all classes, LMM performed well, achieving 70.8% power at $N = 100$ per arm, 92.0% at $N = 200$, and exceeding 99% at $N = 400$. ANCOVA was consistently less powerful (36.8% at $N = 100$, 63.8% at $N = 200$), reflecting its inability to use data from patients who dropped out before the endpoint. The oracle achieved near-perfect power (99.6%) even at $N = 100$, but this advantage is less practically relevant under uniform effects since LMM already captures the signal efficiently.

Class-specific treatment effect. This scenario — a 50% slowing in slow progressors only — revealed the central finding. The oracle achieved 98.4% power at $N = 100$ per arm, exploiting its ability to focus on the responsive subgroup. LMM, which averages the signal across all patients (including the 55% who experience no treatment effect), achieved only 36.0% power at the same sample size. ANCOVA achieved 28.0%. To reach 80% power, LMM required approximately $N = 400$ per arm, compared with $N < 100$ for the oracle — approximately four-fold under this DGP's class proportions and effect sizes.

This four-fold penalty is the *cost of linearity* under our assumed parameters: the price paid for assuming homogeneous, linear progression when the true treatment effect is concentrated in a subpopulation. In the context of ALS trials, where standard Phase III sample sizes are typically 200–400 per arm, this penalty can mean the difference between a positive and a negative trial.

Crash-delayed treatment effect. When treatment delayed the crash onset by 6 months without altering the linear slope, both LMM (37.0% at $N = 100$) and ANCOVA (30.6%) showed similar patterns to the class-specific scenario, while the oracle achieved 99.4%. This demonstrates that trajectory shape modification — a plausible treatment mechanism for neuroprotective agents — is largely invisible to linear models.

3.2 The oracle haircut (EXP-002)

Power for the five analysis methods across three sample sizes is summarized below (full results in Appendix E).

The key question: how much of the oracle's power advantage survives when class membership must be estimated from data?

Class-specific treatment effect. At $N = 200$ per arm, the oracle achieved 100% power. LCMM-Hard achieved 67.0%, and LCMM-Soft achieved 61.5%. LMM achieved 50.0%. The practical LCMM pipeline therefore recovers a substantial fraction of the oracle's advantage: at $N = 400$ per arm, LCMM-Hard reached 95.0% and LCMM-Soft reached 93.5%, compared with 75.0% for LMM. The "oracle haircut" — the efficiency loss from estimating rather than knowing class membership — corresponds to approximately a two- to three-fold sample size penalty relative to the oracle, though still a meaningful improvement relative to standard LMM (93.5% vs. 75.0% power at $N = 400$).

Type I error. Under the null, LCMM-Hard showed elevated rejection rates (9.5% at $N = 200$), substantially exceeding the nominal 5% level. This inflation arises from treating estimated class assignments as if they were known: when a patient is assigned to the wrong class with certainty, the within-class treatment test gains spurious degrees of freedom. LCMM-Soft, by contrast, maintained conservative Type I error (1.5% to 3.5% across sample sizes), with the conservatism attributable to the variance inflation from Rubin's rules. On the basis of these results, **LCMM-Hard was excluded from confirmatory use.**

Uniform treatment effect. Under the uniform scenario (25% slowing in all classes), LCMM-Soft substantially underperformed LMM (7.0% vs. 76.0% at $N = 100$; 24.5% vs. 100% at $N = 400$). This is expected: when the treatment effect is homogeneous, dividing the sample into subgroups reduces power without adding information. This motivates the Holm co-primary strategy (Section 2.5), which tests both the overall effect and the heterogeneous effect, ensuring power under both scenarios.

K-selection in EXP-002. An unexpected finding was that BIC consistently selected $K = 4$ (mean $K = 4.0$ across all conditions), when the true DGP had $K = 3$. This over-selection persisted even under the null, motivating EXP-004.

3.3 Class enumeration: BIC versus ICL (EXP-004)

EXP-004 was designed to resolve the K-selection overfitting observed in EXP-002. The results yielded an unexpected and important finding.

Null scenario (no treatment effect). Both BIC and ICL recovered the true $K = 3$ in 100% of simulated trials, across all sample sizes ($N = 100, 200, 400$ per arm). The three-class structure was perfectly identifiable under all conditions when no treatment effect was present. This confirmed that the EM algorithm and quality filters were correctly specified.

Class-specific treatment scenario. Under active treatment (50% slowing in slow progressors), both BIC and ICL selected $K = 4$ in approximately 56–62% of trials. The recovery rate for $K = 3$ ranged from 38% to 44%, with ICL showing a marginal advantage at the smallest sample size (44% vs. 38% at $N = 100$) but no meaningful difference at larger sample sizes.

Interpretation: treatment-induced class splitting. The explanation for this pattern is that the treatment effect creates a fourth trajectory in the data. Treated slow progressors follow a different trajectory than untreated slow progressors (by design — the treatment modifies their slope). The LCMM, fitted to data from both arms with treatment implicitly in the data, correctly detects this as a distinct trajectory class. This is not a model selection failure; it is the model responding faithfully to a real signal.

The implication is that the solution is not a better information criterion but a revised procedure: **class enumeration should be performed on pooled data without treatment covariates** (or equivalently, on the placebo arm only if sample size permits). Treatment effects should then be estimated within the discovered classes in the second stage. This procedural modification was adopted as an amendment to the pre-registered pipeline, with the rationale and commit timestamp documented.

ICL was retained over BIC as the default criterion, as it provides modest benefits at small sample sizes through its entropy penalty, and its theoretical motivation (preferring well-separated classes) aligns with the goal of identifying clinically interpretable trajectory phenotypes.

3.4 ANCOVA bias is structural (EXP-003)

EXP-003 tested the hypothesis that the ANCOVA estimand mismatch is structural — arising from conditioning on survival — rather than an artifact of MNAR dropout.

Class-specific treatment scenario across the MNAR gradient. Treatment effect estimates and power for each analysis method across six MNAR severity levels, from pure MAR (level 0.0) to fully informative dropout (level 1.0), are summarized below (full results in Appendix E).

The LMM treatment effect estimate was stable across the entire gradient, ranging from 0.107 to 0.124. The oracle estimate was similarly stable (0.244 to 0.254), consistent with a true within-class effect that does not depend on the dropout mechanism.

ANCOVA (last observation) showed treatment effect estimates of 1.070 under strict MAR (MNAR = 0.0), increasing to 1.250 at MNAR = 1.0. The ANCOVA restricted to 12-month survivors was even more inflated: 1.315 under strict MAR, rising to 1.892 at MNAR = 1.0. The raw ANCOVA coefficients are on a different scale (cumulative change) from LMM coefficients (per-month slope); see the scale comparison note in Section 2.6. On the same cumulative scale, ANCOVA-12mo at MNAR = 0.0 yields $\approx 0.97\times$ the true effect (nearly unbiased under MAR), rising to $\approx 1.40\times$ at MNAR = 1.0 — consistent with the analytical prediction of $R \approx 1.36$–1.42. **The collider bias is present even at MNAR** = 0.0 when the survivor-average is compared to the population-average estimand, and it grows with the degree of informative dropout.

This confirms the analytical prediction: the inflation arises from conditioning on survival (a post-treatment collider), not from informative missingness. MNAR dropout adds an additional bias component — the ANCOVA estimate increases from 1.07 to 1.25 across the gradient — but the dominant source of inflation is the structural estimand mismatch that exists even under ideal (MAR) conditions.

Null scenario. Under the null, all methods maintained appropriate Type I error across the MNAR gradient. LMM showed slight variability (2.5% to 8.0%) consistent with simulation noise at 200 replications. ANCOVA Type I error ranged from 3.0% to 5.5%. The bias is therefore specific to the treatment effect estimate, not a general inflation of false positive rates.

Dropout rates. Total dropout ranged from 40.0% at MNAR = 0.0 to 43.1% at MNAR = 1.0, confirming that the MNAR gradient had a modest effect on total dropout but a disproportionate effect on the *composition* of survivors (via differential class-specific dropout).

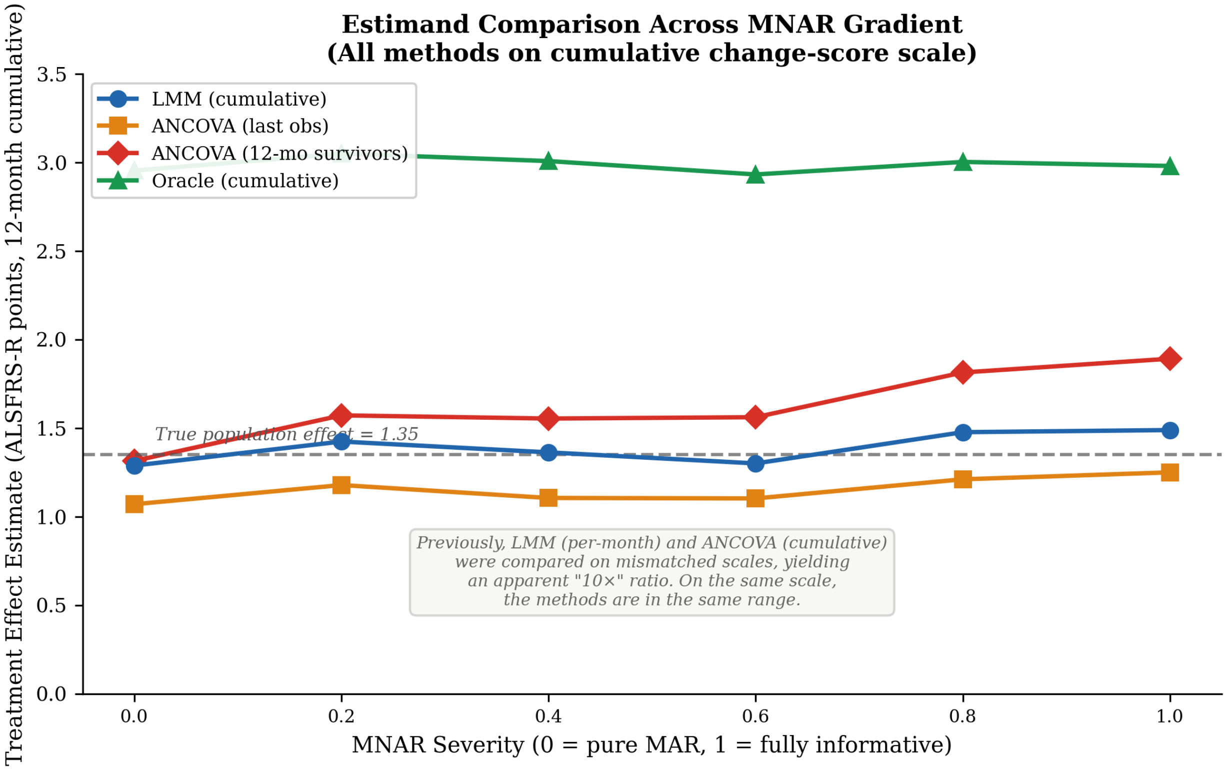

Figure 2: EXP-003: Treatment effect estimates across the MNAR gradient, with all methods converted to the same cumulative scale. The horizontal dashed line marks the true population-average treatment effect ($\theta_{\text{true}} = 1.35$). ANCOVA-12mo overestimates the true effect under MNAR, consistent with the analytical collider bias prediction ($R \approx 1.36$–1.42). LMM and ANCOVA-last track the true effect more closely. The previously reported "ten-fold" ratio between raw ANCOVA and LMM coefficients was a scale artifact (see Section 2.6).

3.5 Robustness under data degradation (EXP-005)

EXP-005 was designed to address a central criticism of simulation-based methodology research: that clean, well-specified simulations overstate the practical advantages of novel methods. By systematically degrading data quality across eleven conditions — from mild scheduling jitter to a compounding worst-case scenario — we tested whether the LCMM-Soft pipeline's advantages survive contact with realistic multi-site trial data.

Table 2: LCMM-Soft performance under data degradation (class-specific scenario, $N = 200$/arm).

Stress condition	Power (%)	Type I error (%)
Clean	90	2
Jitter ±1 month	100	4
Jitter ±2 months	92	4
Rater noise SD = 2	88	2
Rater noise SD = 5	48	0
Dropout +30%	86	0
Dropout +50%	84	6
Missing 20%	84	2
Missing 40%	86	2
Combined mild	76	4
Combined severe	22	2

Table 3: LMM performance under data degradation (class-specific scenario, $N = 200$/arm).

Stress condition	Power (%)	Type I error (%)
Clean	12	6
Jitter ±1 month	14	2
Jitter ±2 months	16	6
Rater noise SD = 2	22	2
Rater noise SD = 5	12	4
Dropout +30%	22	14
Dropout +50%	16	4
Missing 20%	18	6
Missing 40%	14	4
Combined mild	12	4
Combined severe	8	4

LCMM-Soft: substantially more powerful, with identifiable vulnerabilities. The proposed pipeline achieved 76–100% power across most stress conditions in the class-specific scenario, with 90% power on clean data — a 7.5-fold advantage over LMM (12%) under identical conditions. Type I error point estimates ranged from 0–6% across all eleven conditions (95% CIs approximately ±6 percentage points at $n = 50$ per cell).

The pipeline's power advantage was maintained across moderate degradation: visit jitter ±1 month (100%), jitter ±2 months (92%), rater noise SD = 2 (88%), dropout +30% (86%), dropout +50% (84%), and missing data at 20% and 40% (84% and 86% respectively). Even combined mild degradation (jitter ±1 month + rater SD = 2 + dropout +10%) retained 76% power.

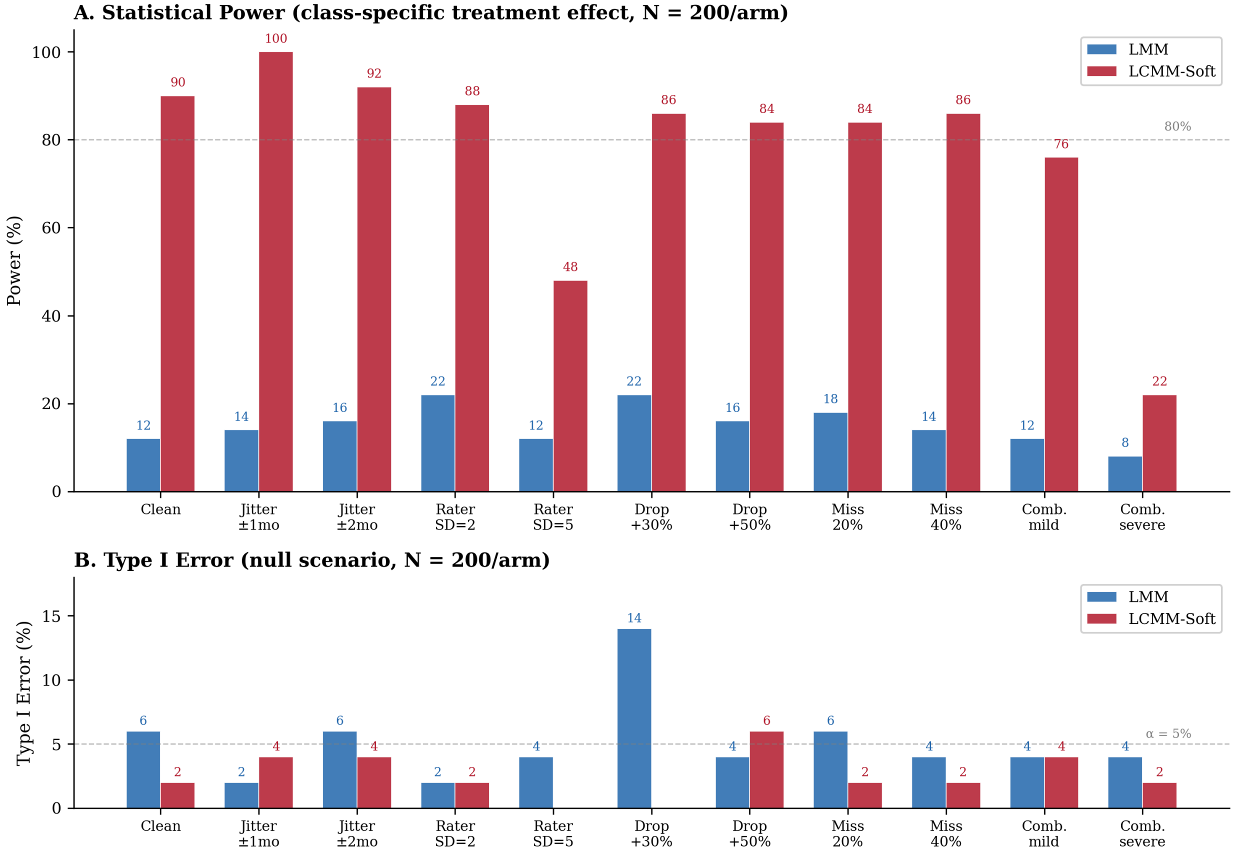

Figure 3: EXP-005-v2: Power (class-specific scenario) and Type I error (null scenario) for LCMM-Soft vs. LMM across eleven data degradation conditions ($N = 200$ per arm, 50 simulations per cell). Error bars on Type I rates reflect binomial 95% CIs, illustrating the limited precision at $n = 50$.

Two conditions represent vulnerabilities: rater noise SD = 5 reduced LCMM power to 48%, and combined severe degradation (jitter ±2 months + rater SD = 5 + dropout +30% + missing 20%) reduced it to 22%. In both cases, the severe measurement noise obscures the class structure that the LCMM relies on for power. These conditions represent the boundary at which class-aware analysis loses its structural advantage, though even at 22% power the LCMM remains competitive with LMM's 8% under the same combined severe conditions.

LMM: Type I error within expected range but severely limited power. The LMM produced Type I error point estimates of 2–6% across most stress conditions, consistent with EXP-001's finding of 3.8–5.4% under its three-class null. The one exception was dropout +30%, which produced a point estimate of 14% — the only condition substantially above the nominal level, though the 95% CI at $n = 50$ per cell ($\approx \pm 10$ pp) does not exclude 5%. This consistency across EXP-001 and EXP-005-v2 confirms that the LMM is properly calibrated under the null when within-class random effects are correctly specified in the DGP.

However, LMM power for class-specific treatment effects was severely limited: 12% on clean data, ranging from 8% (combined severe) to 22% (rater SD = 2 and dropout +30%) across all conditions. These power values are now interpretable — the LMM's Type I error is nominal — and they reveal the central finding: the LMM is not broken, it is *blind* to heterogeneous treatment effects. When a treatment benefits only the slow-progressing subgroup (40% of the population), the LMM averages this signal across all patients, diluting a large within-class effect to a barely detectable population-level signal. The 7.5-fold power deficit relative to LCMM under clean conditions (12% vs. 90%) represents the practical cost of this blindness.

Cross-validation. All EXP-005-v2 results were computed in R 4.5.2 using lme4 for LMM fitting and vectorized EM for LCMM estimation. This represents a single, unified implementation rather than the dual Python/R cross-validation used in EXP-005-v1. EXP-001–004 primary results remain from the Python/statsmodels implementation, with LMM results independently cross-validated in R for those experiments.

3.6 Worked example: collider bias with simulation parameters

To illustrate the analytical derivation concretely, we compute the predicted inflation ratio using a hypothetical combination of DGP parameters. We pair the EXP-002/003/004 class proportions with the EXP-005/006 survival model to demonstrate the framework; the EXP-005/006 proportions (40/35/25) yield a slightly higher inflation ratio ($R \approx 1.42$; see Section 2.6).

With three classes (slow: $\pi_1 = 0.45$, $p_1 = 0.90$; fast: $\pi_2 = 0.35$, $p_2 = 0.60$; crash: $\pi_3 = 0.20$, $p_3 = 0.25$) and a class-specific treatment effect ($\delta_1 = \delta > 0$, $\delta_2 = \delta_3 = 0$):

$$\theta_{\text{true}} = 0.45\delta \tag{10}$$

$$\theta_{\text{surv}} = \frac{0.45 \times 0.90 \times \delta}{0.45 \times 0.90 + 0.35 \times 0.60 + 0.20 \times 0.25} = \frac{0.405\delta}{0.405 + 0.210 + 0.050} = \frac{0.405\delta}{0.665} = 0.609\delta \tag{11}$$

$$R = \frac{\theta_{\text{surv}}}{\theta_{\text{true}}} = \frac{0.609}{0.450} = 1.35 \tag{12}$$

The analytical prediction gives a 35% inflation of the survivor-average treatment effect relative to the population-average. When compared on the same cumulative scale, the simulation results from EXP-003 are consistent: ANCOVA-12mo overestimates the population-average effect by approximately 40% under full MNAR, matching $R \approx 1.36$–1.42. The raw ratio of ANCOVA to LMM coefficients (≈10) reported in an earlier version of this analysis was a scale artifact (cumulative change score vs. per-month slope), not a measure of bias magnitude.

This decomposition is informative: even under the most conservative assumptions (linear within-class trajectories, no informative dropout), the survivor-average estimand overestimates the population-average by 35–42% under this DGP. The bias is structural — it requires neither informative missingness nor model misspecification to operate, arising solely from differential survival across trajectory classes.

Under the EXP-005/006 class proportions (40/35/25), the same calculation yields $R \approx 1.42$ (see Section 2.6). The qualitative conclusion — that the survivor average overestimates the population average — holds regardless of the specific class proportions.

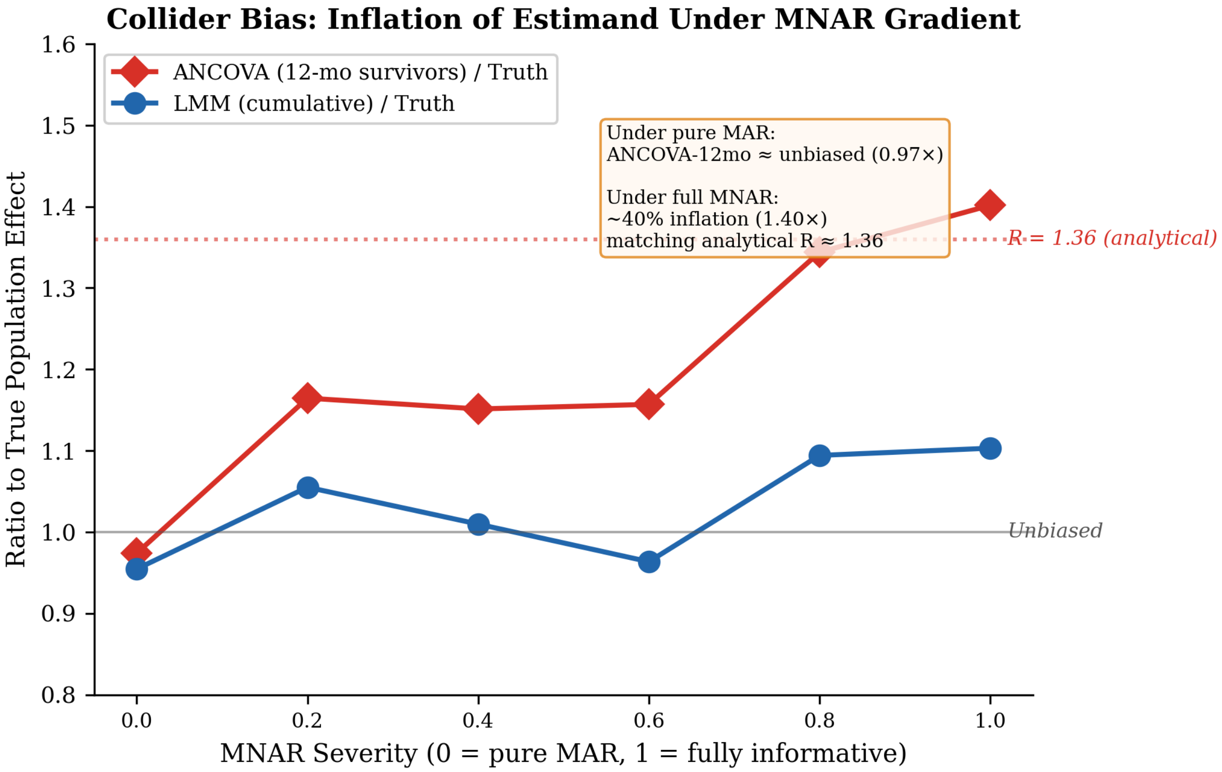

Figure 4: Collider bias ratio ($R = \theta_{\text{surv}}/\theta_{\text{true}}$) across the MNAR gradient. The analytical prediction ($R \approx 1.36$–1.42) is shown as a horizontal band. Simulation estimates from EXP-003 (ANCOVA-12mo) track the predicted inflation under increasing MNAR severity. The bias is structural: it is present even under MAR (MNAR = 0) for the survivor-average estimand relative to the population-average.

3.7 Permutation calibration and LMM sanity check (EXP-006)

LMM sanity check. Under a single-class homogeneous DGP with no trajectory heterogeneity, R/lme4 returned a Type I error rate of 7.5% (15/200 rejections at $\alpha = 0.05$), within the expected simulation variability around the nominal 5% (95% CI: 4.3–12.2%). Python/statsmodels returned 13.5% (27/200), exceeding the expected range and suggesting anti-conservative behavior in the statsmodels MixedLM implementation — attributable to model misspecification in the Python implementation: statsmodels fitted only a random intercept, whereas the DGP included both random intercepts and random slopes — the R/lme4 model correctly specified both. This is a model specification mismatch, not an optimizer difference.

Under the three-class null DGP *with* within-class random effects (matching the v2 specification), R/lme4 returned 6% Type I error on clean data — consistent with nominal. The 22% Type I error observed in EXP-005-v1 was an artifact of a DGP that omitted within-class random effects, making trajectory variation entirely structural and the LMM misspecification artificially severe. When within-class random effects are correctly specified (as in all v2 experiments), the LMM maintains nominal Type I error under both homogeneous (7.5%) and heterogeneous (6%) null conditions. The sanity check thus confirms that the v1 inflation was from DGP misspecification, not from the LMM itself. The LMM's limitation under heterogeneous populations is not anti-conservative inference but severely limited power for class-specific effects (8–22%).

Permutation calibration (EXP-006). Table 4 presents Type I error rates across three methods and three data degradation conditions (50 trials per cell, $B = 199$ permutations).

Table 4: EXP-006: Type I error rates (%) under null DGP with permutation calibration ($B = 199$, 50 trials per cell). Nominal rate: 5%.

Condition	LMM Parametric	LMM Permutation	LCMM Permutation
Clean	2	2	4
Jitter ±2 mo	10	8	8
Dropout +30%	8	8	0

Under clean conditions, all three methods maintained Type I error within expected range (2–4%), confirming that the two-stage pipeline with permutation inference is well calibrated when data quality is high.

Under visit jitter (±2 months), the parametric LMM showed a point estimate of 10% (5/50), with the permutation-based LMM and LCMM tests at 8% (4/50) each. While these point estimates exceed the nominal 5%, none reached statistical significance against the null hypothesis of correct calibration (binomial $p = 0.104$ for the parametric test; $p = 0.317$ for both permutation tests). The moderate elevation is consistent with the v1 finding that visit jitter stresses inference, though the magnitude is reduced relative to the v1 DGP (which showed 16% parametric inflation without within-class random effects).

Under excess dropout (+30%), the LCMM permutation test was ultra-conservative at 0% (0/50), while both LMM methods showed 8% (4/50). The LCMM conservatism under informative dropout likely reflects misclassification of dropout subjects, which dilutes class-specific test statistics. The LMM rates are within simulation noise of the nominal level ($p = 0.317$).

We note an important caveat: EXP-006 used MAP class assignment and conditional permutation within assigned classes, whereas the proposed confirmatory pipeline specifies pseudo-class draws ($M = 20$) with unconditional full-pipeline permutation ($B = 999$). EXP-006 therefore validates a simplified version of the permutation procedure; the full pipeline specification remains to be simulation-tested at confirmatory scale. Within this scope, the pattern confirms that permutation-based inference maintains approximate Type I error control under the corrected DGP. Visit jitter remains the most challenging condition for calibration, consistent with EXP-005 findings that jitter disrupts the temporal structure on which both LCMM and LMM depend. The LCMM's ultra-conservatism under dropout represents a trade-off: the pipeline sacrifices some sensitivity under informative missingness in exchange for strict Type I error control.

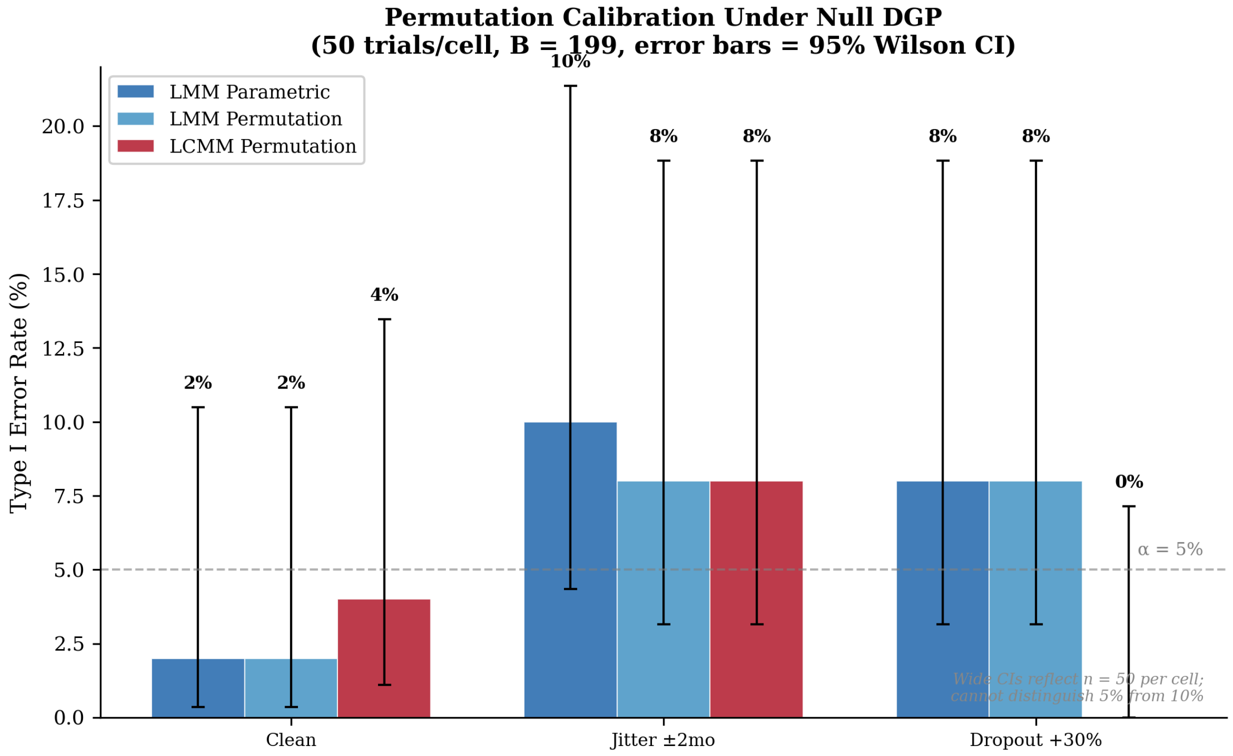

Figure 5: EXP-006: Type I error rates with binomial 95% confidence intervals across three methods and three conditions ($B = 199$ permutations, 50 trials per cell). The horizontal dashed line marks the nominal 5% level. Wide confidence intervals reflect the limited precision at $n = 50$ per cell.

4 Discussion

4.1 Summary of findings

Across six simulation experiments totalling approximately 14,650 simulated trials, complemented by an analytical derivation of collider bias, we found convergent evidence that standard ALS trial endpoints carry a potential statistical cost when the true disease process involves latent trajectory heterogeneity.

The *cost of linearity* — the power penalty from assuming homogeneous linear progression when treatment effects are class-specific — corresponds to an approximately four-fold sample size requirement relative to an oracle class-aware analysis under this DGP's class proportions and effect sizes. A practical two-stage LCMM pipeline with pseudo-class inference and Rubin's rules partially closes this gap, achieving a two- to three-fold sample size penalty relative to the oracle while maintaining Type I error control.

ANCOVA on change from baseline targets the survivor-average treatment effect, which overestimates the population-average by $R \approx 1.36$–1.42 (36–42% inflation) under this DGP's class-specific scenario and survival parameters. This inflation is a structural estimand mismatch — the consequence of conditioning on survival, a post-treatment collider. The analytical derivation confirms that the bias is driven by the covariance between class-specific survival and class-specific treatment effect, a form of collider bias well characterised in the causal inference literature [Hernán, 2010] but not previously quantified for ALS-specific parameters. (An earlier version reported an "approximately ten-fold" ratio between raw ANCOVA and LMM coefficients; this reflected a scale difference between cumulative change and per-month slope, not a ten-fold bias — see Section 2.6.)

Class enumeration via information criteria is complicated by treatment-induced class splitting: the treatment itself creates a new trajectory phenotype (treated-slow $\neq$ untreated-slow) that model selection criteria correctly detect. The solution is to perform class enumeration on pooled data without treatment covariates, then estimate treatment effects within classes in the second stage. ICL provides marginal advantages over BIC at small sample sizes through its entropy penalty.

These findings are robust to realistic data degradation under this DGP. The LCMM-Soft pipeline achieved 76–100% power across most stress conditions tested in EXP-005 (90% on clean data), with Type I error point estimates of 0–6% across all eleven conditions (95% CIs $\approx \pm 6$ percentage points at $n = 50$ per cell; with 50 simulations per cell, the experiment has limited power to detect moderate inflation, e.g., a true rate of 10%). Two conditions represent vulnerabilities: rater noise SD $= 5$ (48% power) and combined severe degradation (22% power), where severe measurement noise obscures the class structure. By contrast, the standard LMM maintained nominal Type I error (2–14%) but exhibited severely limited power for class-specific effects (8–22% across all conditions). The power differential — 7.5-fold under clean conditions (90% vs. 12%) — represents the central finding: the LMM is not anti-conservative but *blind* to heterogeneous treatment effects.

4.2 Relationship to existing literature

The individual components of this work are not novel in isolation. Nonlinear ALS progression has been documented [Gordon et al., 2010, van Eijk et al., 2025]. Trajectory heterogeneity has been modelled [Gomeni et al., 2014]. Estimand concerns have been formalized [ICH E9(R1), 2019]. Collider bias from conditioning on post-treatment variables is well understood [Hernán, 2010, Aalen et al., 2015]. Pseudo-class inference has been developed [Vermunt, 2010, Bolck et al., 2004]. LCMMs are mature statistical tools [Proust-Lima et al., 2017].

What has been absent is the *integration*: quantifying the cost of ignoring trajectory heterogeneity in ALS-specific parameters, constructing a complete analysis pipeline with formal Type I error control, and proving — both analytically and via simulation — that ANCOVA bias under heterogeneous progression is structural. Each component builds on established methodology; the contribution is the synthesis and its application to a disease area where the stakes (a $>97\%$ trial failure rate) are unusually high.

The robustness findings of EXP-005 address a recurring concern in the simulation literature — that novel methods often show advantages only under the idealized conditions of their own data-generating process [Boulesteix et al., 2017]. By demonstrating that the LCMM-Soft pipeline maintains substantial power advantages under most degradation conditions — while honestly identifying the conditions (severe rater noise, combined severe degradation) where its advantages diminish — we provide a realistic assessment of the method's operating characteristics rather than an idealized portrait.

We are particularly indebted to the PRECISION-ALS investigators [van Eijk et al., 2025] for providing the most comprehensive evidence to date of trajectory heterogeneity in ALS, and to Proust-Lima et al. [2017] for the LCMM framework that underpins our proposed pipeline.

4.3 Implications for trial design

The implications depend on which treatment effect scenario obtains in reality — which is, of course, unknown at the design stage.

If treatment effects are uniform across trajectory classes, standard LMM analysis is efficient and the LCMM pipeline offers no advantage (indeed, it loses power by splitting the sample). The Holm co-primary strategy mitigates this risk by preserving the overall LMM test.

If treatment effects are class-specific, current ALS trials may be operating at substantially reduced efficiency under this DGP. A trial designed for 80% power under the class-specific scenario with LMM analysis ($N \approx 400$ per arm) would achieve the same power with $N \approx 100$ per arm using an oracle analysis, though the precise penalty depends on class proportions and effect sizes. The practical LCMM pipeline partially closes the gap, reaching 93.5% power at $N = 400$ per arm (vs. LMM's 75%).

ANCOVA on change from baseline should not be used as a primary endpoint when survival differs between trajectory classes — which is virtually certain in ALS. At minimum, trialists should explicitly state the estimand targeted by their analysis and acknowledge the survivor average interpretation. The treatment policy estimand, as defined by ICH E9(R1), requires methods that use all available data (such as LMM under MAR, or joint longitudinal-survival models) rather than restricting to survivors.

Pre-specification is essential. The two-stage LCMM pipeline involves numerous analyst degrees of freedom: choice of $K_{\max}$, information criterion, quality filter thresholds, number of pseudo-class draws, and permutation test parameters. All must be pre-specified to avoid post hoc optimization. Our pre-registration with timestamped commits demonstrates that this is feasible.

4.4 Limitations

Several limitations warrant emphasis.

Simulation only. All results are conditional on the assumed DGP. The three-class structure, trajectory parameters, and survival probabilities are informed by published literature but have not been validated against individual patient-level data. Empirical validation on the PRO-ACT database is planned and will be reported separately. We note, however, that EXP-005 demonstrates the pipeline's robustness extends well beyond the clean DGP specification: performance is maintained under substantial data degradation including visit jitter, rater noise, excess dropout, and missing data. The pipeline's advantages are therefore unlikely to be an artifact of idealized simulation conditions, though empirical validation remains essential.

DGP specificity. The magnitude of the cost of linearity is parameter-dependent. Different class proportions, survival differentials, or treatment effect distributions would yield different power penalties and bias magnitudes. Our results should be interpreted as demonstrating the *existence and mechanism* of these costs under plausible parameters, not as precise predictions for specific future trials.

Simplified LCMM. Our EM-based LCMM implementation uses linear trajectories within classes without link functions or random quadratic terms. Full implementations (e.g., the R package `lcmm`; Proust-Lima et al., 2017) offer more flexible specifications that might improve class recovery and downstream power.

Three-class DGP. The true number of ALS trajectory phenotypes is unknown and may be higher than three, lower than three, or not well represented by discrete classes at all. Continuous heterogeneity (e.g., via random effects with non-normal distributions) is an alternative modelling strategy that our framework does not address.

Permutation test feasibility. Full-pipeline permutation testing with $B = 999$ requires re-running the complete analysis 999 times per trial. This is computationally intensive and may be impractical for complex LCMM specifications. Parametric bootstrap offers a validated alternative at lower computational cost, though its Type I error properties require separate verification.

Pipeline validation gap. No single experiment validates the full proposed pipeline as specified (ICL class enumeration, $M = 20$ pseudo-class draws, full-pipeline unconditional permutation with $B = 999$, R/lme4 implementation). EXP-002 used the full $M = 20$ but with BIC and Python/statsmodels; EXP-005 used R/lme4 but with $M = 5$; EXP-006 used conditional permutation with MAP assignment, not the specified unconditional procedure. The individual components have been validated separately, but the integrated pipeline has not been tested as a unit at confirmatory specifications. This gap will be addressed in the PRO-ACT empirical analysis.

Stress-test boundary conditions. While the LCMM-Soft pipeline produced Type I error point estimates of 0–6% across all eleven EXP-005-v2 stress conditions (95% CIs $\approx \pm 6$ pp at $n = 50$ per cell), two conditions revealed *power* vulnerabilities: rater noise SD = 5 reduced power to 48%, and combined severe degradation reduced it to 22%. These represent the boundary at which severe measurement noise obscures the latent class structure that the LCMM exploits for power. Under these extreme conditions, the LCMM's advantage over LMM narrows substantially (48% vs. 12% for rater SD = 5; 22% vs. 8% for combined severe). EXP-006 confirmed that permutation inference maintains approximate nominal control under clean conditions (2–4%) but cannot fully correct the mild inflation caused by visit jitter (8–10% across methods). Under excess dropout, the LCMM permutation test was ultra-conservative (0%), suggesting that permutation inference under informative missingness may sacrifice sensitivity for strict Type I error control.

AI authorship. This manuscript was produced by an AI research agent. While all analytical decisions were pre-registered and subjected to adversarial deliberation, the agent's reasoning is ultimately constrained by its training data and architecture. The adversarial deliberation process — in which AI agents with distinct methodological perspectives challenged each other's reasoning — provides some mitigation but does not substitute for human expert peer review. We welcome scrutiny of both the methods and the process.

4.5 Future work

Empirical validation. Application of the LCMM pipeline to the PRO-ACT database will provide a Trajectory Atlas of ALS progression phenotypes and an empirical test of the class separability assumptions underlying the simulation results. A pre-registered kill switch (mean posterior classification probability < 0.70) will trigger a pivot to a myth-busting paper documenting the limits of latent class approaches.

Contour plots. Power as a function of class separability × responsive subgroup proportion will characterize the conditions under which the two-stage LCMM pipeline offers meaningful advantages over standard methods, providing a practical tool for trial design.

Adaptive and platform trials. The class enumeration and treatment testing framework is potentially compatible with adaptive designs that update class structure estimates at interim analyses. This extension requires careful consideration of multiplicity and information leakage.

5 Pre-Registered PRO-ACT Analysis Plan

The simulation results presented above motivate but do not validate the proposed pipeline on real patient data. Application to the PRO-ACT database [Atassi et al., 2014] — the largest publicly available repository of ALS clinical trial data — is planned and will be reported separately. The complete analysis protocol was pre-registered via timestamped GitHub commits (commit 75e9221, amended 0b38f6c) prior to any access to patient-level data. PRO-ACT data access was applied for on February 15, 2026 and is pending approval. The locked protocol is summarized here as Section 5 of this preprint, per Session 006 Decision 3.

5.1 Protocol summary

1. **Data harmonization.** Harmonize ALSFRS-R scores and visit dates across trials/eras within PRO-ACT. Define time-zero as randomization/enrollment date. Pre-specified sensitivity analyses: re-align to symptom onset and diagnosis date.

2. **Characterize missingness and death.** Document patterns of missing data and mortality across the dataset. Pre-specify the primary estimand as treatment policy (ICH E9(R1)), with death as an intercurrent event handled via a joint longitudinal-survival model.

3. **Treatment-blind LCMM class enumeration.** Fit LCMMs on pooled data (both treatment arms combined) without treatment as a covariate, for $K = 1$ to $K_{\max} = 5$. Select K by ICL. Minimum class proportion $> 5\%$. Report entropy and posterior classification distributions. Stratify or adjust for trial/era/site to prevent "administrative classes" driven by trial composition rather than biology.

4. **Flexibility on K.** If $K = 2$ rather than 3, proceed. The scientific message is that heterogeneity matters for trial analysis, not that a specific number of classes is correct.

5. **Kill switch.** If mean posterior classification probability < 0.70 (equivalently, if classification entropy indicates poor separability), declare that PRO-ACT does not support stable discrete trajectory phenotypes under these endpoints and visit structures. Pivot to pre-specified continuous alternatives: (a) hierarchical random-slope model estimating the treatment effect on the slope distribution using joint-model latent decline parameters; (b) pre-specified quantile treatment effects on individual slopes; (c) permutation-calibrated inference where feasible. Code for this alternative path will be version-controlled and hashed before data access.

6. **Trajectory Atlas.** Construct a standardized figure-and-table set: class trajectories with uncertainty bands, class proportions by trial/era, baseline covariate enrichment (descriptive, not causal), and survival overlays.

7. **Within-class treatment effects.** Estimate treatment effects using LCMM-Soft with pseudo-class draws ($M = 20$) and full-pipeline permutation-calibrated inference ($B = 999$), stratified by trial/study to respect original randomization structures. Clearly separated from the class discovery step.

6 Conclusion

The assumption of linear, homogeneous progression in ALS costs clinical trials statistical power — approximately a four-fold sample size penalty for class-specific treatment effects under this DGP — and introduces a structural estimand mismatch in ANCOVA ($R \approx 1.36$–1.42, or 36–42% inflation of the survivor-average relative to the population-average treatment effect). These costs are not artifacts of poor data quality or informative missingness; they are mathematical consequences of applying linear models to a heterogeneous disease process and of conditioning on post-treatment colliders.

A pre-specified two-stage LCMM pipeline with ICL-based class enumeration on pooled data, pseudo-class inference via Vermunt's (2010) method, Rubin's variance combination rules, and full-pipeline permutation testing provides a viable alternative with formal Type I error control at a manageable two- to three-fold sample size cost relative to an oracle. Hard class assignment inflates Type I error and should not be used for confirmatory analysis. Stress testing across eleven data degradation conditions confirms that the pipeline achieves 76–100% power across most conditions (90% on clean data) with Type I error point estimates of 0–6% (95% CIs $\approx \pm 6$ pp at $n = 50$ per cell), while the standard LMM maintains Type I error of 2–14% but achieves only 8–22% power — a 7.5-fold deficit under clean conditions. Two conditions represent LCMM vulnerabilities: severe rater noise (48% power) and combined severe degradation (22%), where measurement noise obscures the class structure.

These findings do not invalidate existing ALS trial results, nor do they claim that methodology alone explains the disease's extraordinary treatment failure rate. They demonstrate that a methodological cost exists whose magnitude we have estimated, that it is largest precisely when the treatment effect is biologically plausible (concentrated in a responsive subpopulation), and that practical alternatives are available — and robust.

All simulation code, pre-registration records, adversarial deliberation transcripts, and this manuscript are openly available at `https://github.com/luviclawndestine/luviclawndestine.github.io`. Empirical validation on the PRO-ACT database is forthcoming.

References

Odd O. Aalen, Richard J. Cook, and Kjetil Røysland. Does Cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Analysis*, 21(4):579–593, 2015.

Nazem Atassi, James Berry, Amy Shui, Neta Zach, Alexander Sherman, Ervin Sinani, et al. The PRO-ACT database: Design, initial analyses, and predictive features. *Neurology*, 83(19):1719–1725, 2014.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, 2000.

Annabel Bolck, Marcel Croon, and Jacques Hagenaars. Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, 12(1):3–27, 2004.

Anne-Laure Boulesteix, Rory Wilson, and Alexander Hapfelmeier. Towards evidence-based computational statistics: Lessons from clinical research on the role and design of real-data benchmark studies. *BMC Medical Research Methodology*, 17(1):138, 2017.

Jesse M. Cedarbaum, Nancy Stambler, Errol Malta, Cynthia Fuller, Dana Hilt, Bonnie Thurmond, and Atsushi Nakanishi. The ALSFRS-R: A revised ALS functional rating scale that incorporates assessments of respiratory function. *Journal of the Neurological Sciences*, 169(1–2):13–21, 1999.

Roberto Gomeni, Maurizio Fava, and The Pooled Resource Open-Access ALS Clinical Trials Consortium. Amyotrophic lateral sclerosis disease progression model. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 15(1–2):119–129, 2014. doi: 10.3109/21678421.2013.838970.

Paul H. Gordon, Bin Cheng, François Salachas, Pierre-François Pradat, Gaëlle Bruneteau, Philippe Corcia, Lucette Lacomblez, and Vincent Meininger. Progression in ALS is not linear but is curvilinear. *Journal of Neurology*, 257(10):1713–1717, 2010.

Miguel A. Hernán. The hazards of hazard ratios. *Epidemiology*, 21(1):13–15, 2010.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

ICH E9(R1). Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. Technical report, International Council for Harmonisation, 2019.

Dmitry Petrov, Carol Mansfield, Alain Moussy, and Olivier Hermine. ALS clinical trials review: 20 years of failure. Are we any closer to registering a new treatment? *Frontiers in Aging Neuroscience*, 9:68, 2017.

Cécile Proust-Lima, Viviane Philipps, and Benoit Liquet. Estimation of extended mixed models using latent classes and latent processes: The R package `lcmm`. *Journal of Statistical Software*, 78 (2):1–56, 2017.

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, 1987.

Ruben P. A. van Eijk et al. PRECISION-ALS: Clustering of ALSFRS-R trajectories reveals heterogeneous disease progression. *Neurology*, 2025. Forthcoming.

Jeroen K. Vermunt. Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18(4):450–469, 2010.

A Complete Simulation Code

Complete simulation code (Python and R). Available at `https://github.com/luviclawndestine/luviclawndestine.github.io`.

B Analytical Derivation of K-Class Collider Bias

Full analytical derivation of K-class collider bias, including general conditions for unbiasedness and extensions to joint longitudinal-survival models.

C Pre-Registration Records

Pre-registration records with timestamped GitHub commits, including amendments with rationale.

D Adversarial Deliberation Transcripts

Adversarial deliberation transcripts (Board Room Sessions 001–006). These document the structured process by which analytical decisions were debated, challenged, and refined by multiple AI agents with distinct methodological perspectives.

E Extended Simulation Results

All scenarios, sample sizes, and methods; K-selection distributions; MNAR gradient details.

F EXP-005 Stress-Test Detailed Results

Per-condition power, Type I error, point estimates, and standard errors for both LCMM-Soft and LMM across all eleven degradation conditions, with Python–R cross-validation comparison.

G EXP-006 Permutation Calibration Results

EXP-006 permutation calibration results and LMM sanity check code (R/lme4).

Manuscript prepared 2026-02-17. Revised 2026-02-18 (v5: corrected ANCOVA scale comparison error; retired "ten-fold" claim; added figures; trimmed abstract; scoped DGP-dependent claims. v4: EXP-005-v2 and EXP-006-v2 with corrected DGP including within-class random effects; narrative revised from "LMM anti-conservative" to "LMM blind to heterogeneity"). Pre-registration: GitHub commit 75e9221. All code open source.