The New England Journal of Statistics in Data Science Volume 4, 170–191 (2026)
DOI: https://doi.org/10.51387/26-NEJSDS96

Modeling the Mean with Time as a Categorical Variable in Longitudinal Designs for Smaller-Sized Clinical Trials: A Case Studies Approach Based on Three Phase 3 Clinical Trials in Rare Diseases

David ZAHRIEH*, Yi WANG, Jennifer LE-RADEMACHER[1], and Tony KOUTSOUKOS[1]

Abstract

Background: Generalized estimating equations (GEE) and mixed-model repeated measures (MMRM) can handle longitudinal continuous outcomes when modeling the mean with time included categorically. Due to small sample sizes in rare diseases, a compound symmetry (CS) covariance pattern is sometimes adopted. In this setting, there is scant literature in the rare disease community that provide practical advice about the use of both methods based on real datasets from trials conducted in rare diseases, including when to use the sandwich variance estimator with or without a bias correction.

Methods: To fill this gap, we simulated data from three longitudinal, phase 3 trials conducted in rare diseases to jointly review the operating characteristics: a randomized trial in GNE myopathy (N = 44 placebo; N = 45 treatment) and pediatric X-linked hypophosphatemia (XLH) (N = 32 control; N = 29 treatment), and a single-arm in adult XLH (N = 14).

Results: In each trial, few participants discontinued; furthermore, <1.5% of the measurement occasions were missing outcome data, no missing outcome data pattern occurred in >1 participant, and the missing completely at random (MCAR) assumption was clinically justified. In the two trials with nonconstant variances/covariances over time, bias-corrected sandwich variance estimators with t-based inference was needed with MMRM and GEE. If the CS pattern was a good approximation, as seen in the pediatric XLH trial, then model-based standard errors with t-based inference performed well with both methods

Conclusion: Based on a review of three case studies, the MCAR assumption was plausible and missingness low. When modeling the mean response with time included categorically and with a parsimonious CS covariance structure, each method required careful consideration with its use.

KEYWORDS AND PHRASES: Generalized estimating equations, Longitudinal designs, Mixed model repeated measures, Rare diseases, Sandwich variance estimator, Small-sample correction.

1. INTRODUCTION

In longitudinal clinical trials the number and the collection times of the repeated measurements are oftentimes the same for all participants. Consequently, there is usually enough replications at each measurement occasion and within each treatment group, in the case of moderately sized or large randomized trials, to estimate the true underlying covariance matrix. Mixed-model repeated measures (MMRM) – a likelihood-based regression model for modeling the mean with time included categorically – has been generally recommended for the primary analysis of continuous outcomes with longitudinal trial designs with the use of unstructured (UN) modeling of within-patient covariances [1]. With the focus on modeling the mean response with time included categorically, the corresponding generalized estimating equations (GEE) method can provide an alternative (i.e., marginal models with an identity link). When the longitudinal measurements are multivariate Gaussian, the solution to the estimating equations is the generalized least squares estimator of the regression coefficients from MMRM [2]. Like MMRM, GEE can handle missing data easily, albeit provided that the longitudinal outcomes satisfy the missing completely at random (MCAR) assumption [3].

A compelling argument can sometimes be made to support the MCAR assumption in well-conducted trials in rare diseases. Patient centeredness is central to trial design. Involving patients, patient advocates, and family members in trial design can increase retention and ensure adequate assessment of trial outcomes [4, 5]. Compared with non-rare disease trials, trials in rare diseases tend to enroll fewer pa-

*Corresponding author.
[1]Contributed equally.

tients, address a high unmet medical need, and use patient-centered or surrogate endpoints to assess efficacy [6, 7]. Such design characteristics can mitigate dropouts and incomplete data. The use of external controls, shorter placebo-phase duration, and designs that assure participants that they will receive the experimental treatment favor patient participation and retention and further reduce the risk of dropout [8, 9]. However, if MCAR is not clinically defensible and dropout is assumed to be missing at random (MAR), the estimating equations need to be weighted by the propensity to dropout [10]. Whereas it is generally appreciated that MMRM provides a valid framework under the MAR assumption, it is not well recognized that this assumption does not hold if the covariance pattern has not been correctly specified [11]; and, therefore, may lead to a false sense of security with its use.

Little priori information about the covariance pattern is available in rare diseases. Furthermore, there are no reliable ways to identify the best covariance model in small samples [12]. If the data allow it, an UN covariance pattern should be considered; it minimizes the risk of a misspecified covariance structure. If convergence problems in numerical optimization are encountered, strategies to achieve convergence should be considered before adopting a parsimonious covariance pattern [1, 13, 14]. In some controlled or single-arm trials in rare diseases, however, there simply are not enough participants relative to the number of time points to estimate the covariance terms with an UN covariance pattern. In either case, and because we regard the correlation structure as a nuisance, the compound symmetry (CS) covariance pattern could pragmatically be considered in trials in rare diseases, which recognizes the dependence at the cost of only one extra covariance parameter.

Modeling time categorically imposes no structure on the mean response over time. It is unlikely, therefore, that the model for the mean response is misspecified; MMRM and GEE estimates of the regression coefficients will be valid. However, because the CS covariance pattern will likely be misspecified, the model-based standard errors (SEs) for MMRM and GEE estimates of the regression coefficients will be invalid [11, 15]. This is true in small and large samples [16]. With both methods, valid SEs of the regression coefficients can be based on the sandwich variance estimator, which are robust to any misspecification of the covariance [15]. Of course, the sandwich variance estimator is based on large-sample properties and its performance deteriorates for small samples [17–19]. Small-sample corrections for these SEs are needed to correct for bias.

Researchers have developed and compared a variety of bias-corrected SEs across a range of simulated scenarios, albeit separately for MMRM and GEE [18–27]. Furthermore, as opposed to relying on large-sampling properties of the regression coefficients when building confidence intervals (CIs) and testing hypotheses for the regression coefficients, $t-$ and F-distributions using degrees of freedom (DOF) approximations to account for uncertainty in estimation of the regression coefficients can be used. Because it is not always clear what defines a small sample, a suitable bias-corrected strategy should be considered for trials conducted in rare diseases along with the choice to use either MMRM or GEE.

While there are numerous statistical publications on using these methods spanning decades, there is a dearth of literature in the rare disease community that uses real data to collectively report on both methods applied to continuous outcomes collected longitudinally from trials in rare diseases that is succinctly presented in an accessible manner. Analysts in rare disease research may be confused on which method to apply in practice and how to use the chosen method correctly. Coupled with the need to specify the primary analysis method in a protocol before the trial begins [28], a case studies approach to prespecifying a modeling strategy in rare diseases was taken. Herein, we considered three unique, albeit representative longitudinal trial designs with continuous efficacy outcomes conducted in rare diseases. We report the degree of missing data, evaluate the missing data pattern, and assess common characteristics of the empirical covariance pattern of the repeated measurements. We then simulated data directly from these trials to share with other analysts in the rare disease community a concise review of the operating characteristics of these methods based on real data. We provide a succinct summary and practical advice to complement the statistical literature on when and how to apply these methods in practice.

2. METHODS

2.1 Efficacy Outcomes

The primary (or important secondary) efficacy outcome from three completed, longitudinal, phase 3 FDA registration trials conducted in rare diseases were selected (Table 1): changes in the HHD upper extremity composite score, measured at baseline and every 8 weeks through study week 48 in UX001-CL301, a phase 3 randomized study evaluating sialic acid extended-release for GNE myopathy (N = 44 placebo; N = 45 treatment) [29]; changes in rickets severity, assessed by the Radiographic Global Impression of Change (RGI-C) global score and measured at weeks 40 and 64 in UX023-CL301, a randomized phase 3 trial (N = 32 control; N = 29 treatment) comparing the efficacy and safety of burosumab to conventional therapy for XLH [30]; and changes in alkaline phosphatase (ALP), measured at baseline and at weeks 4, 12, 24, 36, and 48 in UX023-CL304, a phase 3 single arm study in N = 14 participants investigating the effects of burosumab on osteomalacia in adults with X-linked hypophosphatemia (XLH) [31]. For each outcome, we report the degree and pattern of missing data, evaluated the MCAR assumption using Little's MCAR Test [32], and assessed the empirical covariance pattern of the changes from baseline.

Table 1. The three longitudinal, phase 3 clinical trials conducted in rare diseases, which formed the basis for the simulation study.

Clinical Trial	No. Post-BL Occasions	Trial Duration	Sample Size		Primary (or Important Secondary) Efficacy Endpoint	Prospectively Defined Primary Analysis	ClinTrials.Gov
			Randomized or Enrolled	Primary Analysis Set			
UX001-CL301	6	48 Weeks	89	88	Primary endpoint: Change from baseline at Week 48 in the HHD Upper Extremity (UE) Composite Score	GEE, Change from Baseline, compound symmetry, adjusting for sex, region, and baseline value	NCT02377921
UX023-CL301 (Pediatrics)	2	64 Weeks	61	61	Primary endpoint: Rickets severity, assessed by the Radiographic Global Impression of Change (RGI-C) global score and measured at Week 40	GEE, Change from Baseline, compound symmetry, adjusting for age and the baseline total rickets severity score	NCT02915705
UX023-CL304 (Adults)	5	48 Weeks	14	14	Important secondary endpoint: Change from baseline at Week 48 in alkaline phosphatase (ALP)	GEE, Change from Baseline, compound symmetry, adjusting for the baseline value	NCT02537431

BL = Baseline; GEE = generalized estimating equations.
Trials UX001-CL301 and UX023-CL301 were randomized controlled trials, while UX023-CL304 was a single arm trial.
The prespecified *Primary Analysis Set* was defined as all randomized (or enrolled) participants with a baseline measurement and at least one post-baseline measurement.
Standard GEE was applied such that estimation was based on method-of-moments assuming asymptotic normality using the SAS GENMOD procedure.

The prespecified GEE method, which was the primary analysis method and specified in the protocol prior to seeing the data, modeled the longitudinal changes from baseline assuming a CS covariance pattern adjusting for the corresponding baseline values (baseline total rickets severity score in UX023-CL301) and other prespecified potential confounding factors; the sandwich variance estimator of the regression coefficients without a small-sample correction was used and also prespecified. The SAS procedure GENMOD was used such that estimation was based on the method-of-moments and CIs and tests assumed asymptotic normality [so-called standard GEE; Liang and Zeger (1986)] [3]. In each trial, the prespecified GEE method was applied to the primary analysis set, defined in the protocol as all randomized (or enrolled) participants who had a baseline measurement and $\geq$1 post-baseline measurement and was used to assess a treatment effect at a single time point. The number of participants in each primary analysis set was 88 (43 placebo; 45 treatment), 61 (32 control; 29 treatment), and 14 in UX001-CL301, UX023-CL301, and UX023-CL304, respectively (Table 1).

2.2 Simulation Study

The aim was to understand the operating characteristics of the prespecified GEE method to estimation and MMRM applied to each trial, and to recommend when to use the sandwich variance estimator with or without a small-sample correction with both analytic approaches. Multivariate Gaussian longitudinal outcomes of changes from baseline were generated with a fixed covariance. For a linear model with a given covariance matrix, MMRM and GEE (marginal model with an identify link function) will give identical results of the regression coefficients; however, the methods differ in how they estimate the covariance [14]. Because GEE does not use likelihood methods, the estimated "model" is incomplete and not suitable for simulation. To study the operating characteristics of both methods, data were generated from an estimated MMRM fit to each dataset's outcome (Appendix Table 1). For the two-arm trials (UX001-CL301 and UX023-CL301), the parameterization of the mean part of the model included an intercept, indicators for time, a group indicator, the two-way interaction between the indicators for time and the group indicator, the baseline value and a set of protocol-specified covariates. For the single arm trial (UX023-CL304), the mean part of the model comprised an intercept, indicators for time, and the baseline value. The baseline values and the protocol-specified covariates were generated from a distribution consistent with what was seen in the data. The models and primary statistical hypotheses expressed in mathematical terms are shown in Table 2. Based on re-analysis of the trials' outcomes, we successfully estimated an UN covariance pattern. Compared with CS, the Akaike Information Criterion (AIC) [33] expressed a strong preference [34] (difference in AIC>10) for an UN covariance pattern in UX001-CL301 and UX023-CL304. The lack of a preference (difference in AIC<2) for an UN covariance pattern in UX023-CL301 notwithstanding, a fixed UN covariance pattern was used in data generation for the longitudinal outcomes from UX001-CL301, UX023-CL304, and UX023-CL301. To generate data that were MCAR, we removed each datum independently of the other data with probability 1.4%, 0.5%, and 1.2% in datasets simulated from UX001-CL301, UX023-CL301, and UX023-CL304, respectively. These missing data percentages were based on the real trial data; more dialogue is provided on this subject in the results section.

Table 3 lists the modeling scenarios studied. MMRM was estimated using restricted maximum likelihood with the

Table 2. Data generating model for UX001-CL301, UX023-CL301, and UX023-CL304.

Clinical Trial	MMRM Data Generating Model
UX001-CL301	$E(Y_i \mid X_i) = \beta_0 + \beta_1 I(\text{G} = 1) + \beta_2 I(\text{W} = 16) + \beta_3 I(\text{W} = 24) + \beta_4 I(\text{W} = 32) + \beta_5 I(\text{W} = 40) + \beta_6 I(\text{W} = 48) + \beta_7 I(\text{G} = 1) \times I(\text{W} = 16) + \beta_8 I(\text{G} = 1) \times I(\text{W} = 24) + \beta_9 I(\text{G} = 1) \times I(\text{W} = 32) + \beta_{10} I(\text{G} = 1) \times I(\text{W} = 40) + \beta_{11} I(\text{G} = 1) \times I(\text{W} = 48) + \gamma_1 \text{Base} + \gamma_2 I(\text{Sex} = \text{Male}) + \gamma_3 I(\text{Region} = \text{NonUS})$ Repeated measurements taken on $i = 1, \ldots, 88$ participants (45 treatment; 43 placebo) at the same set of 6 post-baseline occasions. Y_i = Change from baseline. I represents an indicator variable. G = Group; W = Week; Base = Baseline value; Reference categories for variables G, W, Sex, and Region were placebo, week 8, female, and US, respectively. Unstructured (UN) covariance structure was assumed. The estimates of the regression coefficients and variance / covariance matrix from fitting the MMRM model to the trial data were used as the true values. The covariates Base, Sex, and Region were generated from a distribution reasonably consistent with the trial data: a truncated $N(56.14, 27.954)$, 55% males, and 66% Non-US. 100,000 data sets generated; on average, in any given simulated data set, 1.4% of the data were missing completely at random. The regression coefficients of primary interest are β_1 and β_{11}; specifically, the null hypothesis $H_0: \beta_1 + \beta_{11} = 0$ corresponds to no treatment difference in the mean change from baseline at week 48 adjusting for the other covariates; this is similar to testing $H_0:$ Treament Difference in LS Means at Week 48 = 0 (i.e., the marginal mean change from baseline is the same in each arm).
UX023-CL301 (Pediatrics)	$E(Y_i \mid X_i) = \beta_0 + \beta_1 I(\text{G} = 2) + \beta_2 I(\text{W} = 64) + \beta_3 I(\text{G} = 2) \times I(\text{W} = 64) + \gamma_1 \text{TOTSCORE} + \gamma_2 I(\text{Age} = 2)$ Repeated measurements taken on $i = 1, \ldots, 61$ participants (32 conventional therapy; 29 burosumab) at the same set of 2 post-baseline occasions. Y_i = Change from baseline. I represents an indicator variable. G = Group; W = Week; TOTSCORE = Baseline Total Rickets Severity Score; Reference categories for variables G, W, and Age were burosumab, week 40, and < 5 years, respectively. Unstructured (UN) covariance structure was assumed. The estimates of the regression coefficients and variance / covariance matrix from fitting the MMRM model to the trial data were used as the true values. The covariates TOTSCORE and Age were generated from a distribution reasonably consistent with the trial data: a truncated $N(3.18, 1.057)$ and 57% $\geq$ 5 years old. 100,000 data sets were generated; on average, in any given simulated data set, 0.5% of the data were missing completely at random. The regression coefficient of primary interest, β_1, corresponds to the treatment difference in the mean change from baseline at week 40 adjusting for the other covariates; this is similar to testing $H_0:$ Treament Difference in LS Means at Week 40 = 0 (i.e., the marginal mean change from baseline is the same in each arm).
UX023-CL304 (Adults)	$E(Y_i \mid X_i) = \beta_0 + \beta_1 I(\text{W} = 12) + \beta_2 I(\text{W} = 24) + \beta_3 I(\text{W} = 36) + \beta_4 I(\text{W} = 48) + \gamma_1 \text{Base}$ Repeated measurements taken on $i = 1, \ldots, 14$ participants at the same set of 5 post-baseline occasions. Y_i = Change from baseline in alkaline phosphatase. I represents an indicator variable. W = Week; Base = Baseline value; Reference category for the variable W was study week 4. Unstructured (UN) covariance structure was assumed. The estimates of the regression coefficients and variance / covariance matrix from fitting the MMRM model to the trial data were used as the true values. The covariate Base was generated from a distribution reasonably consistent with the trial data: a truncated $N(113.14, 51.600)$. 500,000 data sets were generated; on average, in any given simulated data set, 1.2% of the data were missing completely at random. The regression coefficients of primary interest are β_0 and β_4; specifically, the null hypothesis $H_0: \beta_0 + \beta_4 = 0$ corresponds to no difference in the mean change from baseline at week 48 adjusting for the baseline value. This is similar to testing $H_0:$ LS Mean at Week 48 = 0 (i.e., the marginal mean change from baseline equals zero).

Table 3. For each dataset generated in the simulation study, the following models were applied.

Clinical Trial	Model	Method	Covariance Pattern	Standard Error Adjustment	Small-Sample Correction to Standard Errors	Degree of Freedom (DOF) in t Tests and t-based Confidence Limits
UX001-CL301 & UX023-CL301 (Pediatrics) & UX023-CL304 (Adults)	M0	MMRM	UN	Kenward-Roger	No	Kenward-Roger approximation for the DOF based on the SE adjustment
	M1	MMRM	CS	None, Model-based	No	Between-Within method DOF by Schluchter and Elashoff (1990)
	M2	MMRM	CS	Sandwich Estimator	No	
	M3	MMRM	CS	Sandwich Estimator	Mancl and DeRouen (2001)	
	M4	GEE	CS	None, Model-based	No	Asymptotic normal distribution (standard GEE) with z-based inference
	M5*	GEE	CS	Sandwich Estimator	No	
	M6	GEE	CS	Sandwich Estimator	Mancl and DeRouen (2001)	
	M7	GEE	CS	None, Model-based	No	Between-Within method DOF by Schluchter and Elashoff (1990)
	M8	GEE	CS	Sandwich Estimator	No	
	M9	GEE	CS	Sandwich Estimator	Mancl and DeRouen (2001)	

GEE = generalized estimating equations; MMRM = mixed model repeated measures; SE = standard error; DOF = degrees of freedom; UN = Unstructured; CS = Compound symmetry.

* Prespecified GEE method, which was the primary analysis method defined in the protocol for each trial.

Note 1. Trials UX001-CL301 (N = 88) and UX023-CL301 (N = 61) were randomized controlled trials, while UX023-CL304 (N = 14) was a single arm trial.

Note 2. The SAS procedure PROC GLIMMIX with METHOD = RSPL was used to fit the MMRM models. The SAS Keyword EMPIRICAL = FIRORES was used to obtain the bias-corrected SEs by Mancl and DeRouen (2001) made to the empirical/robust standard errors. The SAS Keyword DDFM = BETWITHIN was used to obtain the between-within DOF approximation by Schluchter and Elashoff (1990): The DOF approximation used for the between and within-subject effects were $N_1 - p_1$ and $N_2 - (N_1 + p_2)$, respectively; here, N_1 = number of analyzable subjects; N_2 = number of nonmissing observations; p_1 = number of between-subject effects + intercept; p_2 = number of within-subject effects. For hypothesis tests that involved within-subject effects, $N_2 - (N_1 + p_2)$ DOF was used.

Note 3. The standard GEE method of moments approach was used that mirrors the SAS procedure PROC GENMOD, albeit with the use of the *GEE WITH DIAGNOSTICS* (Version 1.06) SAS/IML macro by John Preisser, University of North Carolina-Chapel Hill; the SAS/IML macro calculates the model-based, empirical, and bias-corrected SEs (Mancl and DeRouen, 2001).

Note 4. All models adjusted for the prespecified covariates.

SAS procedure GLIMMIX. MMRM with unstructured covariance (Model M0) served as the benchmark. Here, the Kenward-Roger method [35] was considered, which combines a SE adjustment with a Satterthwaite-type DOF approximation. Models M1, M2, and M3 represented an MMRM with CS covariance pattern that, respectively, relied on model-based, sandwich-based, and bias-corrected sandwich-based standard errors by Mancl and DeRouen (2001) [19] by specifying the EMPIRICAL = FIRORES option in the GLIMMIX statement. For models M1-M3, inferences were based on the t-distribution using the between-within method [36] (DDFM = BETWITHIN) for computing DOF. The DOF approximation used for the between- and within-subject effects in the model were $N_1 - p_1$ and $N_2 - (N_1 + p_2)$, respectively, where N_1 = number of analyzable subjects; N_2 = number of nonmissing observations; p_1 = number of between-subject regression coefficients + intercept; p_2 = number of within-subject regression coefficients. For hypothesis tests that involved within-subject effects, $N_2-(N_1+p_2)$ DOF was used. For GEE, estimation was based on method-of-moments assuming a CS covariance pattern (Models M4-M9). Because built-in bias-corrected SEs are not available with GENMOD, we used the SAS/IML GEE macro (Version-1.06) [37], which mirrors the estimation method in GENMOD, but also calculates bias-corrected SEs by Mancl and DeRouen (2001). Inference was based on asymptotic normality (M4-M6) or the t-distribution using the between-within DOF-approximation obtained with MMRM (M7-M9).

We compared the bias and SEs associated with the regression coefficients of primary interest, the coverage of their corresponding 95% CIs, and the SEs and power of the test expressed in terms of the least square means (estimated marginal means) testing the null hypothesis that the treatment difference of the marginal mean change from baseline at weeks 48 and 40 equaled zero in the randomized trials UX001-CL301 and UX023-CL301, respectively, and the null hypothesis that the marginal mean change from baseline at week 48 equaled zero in the single arm trial UX023-CL304. Testing the primary hypotheses in terms of the estimated marginal means was prespecified in each study.

3. RESULTS

Within each trial, Table 4 shows the proportion of missing data, results from applying Little's MCAR Test, and the empirical covariance pattern of the repeated measurements. Two participants in UX001-CL301 and one participant in trial UX023-CL304 discontinued the trial prematurely, while none of the participants in UX023-CL301 discontinued early.

Table 4. Missing data summary and empirical covariance pattern of the changes from baseline for each outcome according to the phase 3 clinical trial.

Phase 3 Clinical Trial	No. Randomized or Enrolled	No. Post-BL Occasions	No. Patients Discontinuing Early	Percent of BL and Post-BL Measurement Occasions Missing Data	Little's MCAR Test[d]	Empirical Covariance Matrix[e] Changes from Baseline						
							WK8	**WK16**	**WK24**	**WK32**	**WK40**	**WK48**
						WK8	**16.03**	8.14	9.68	14.96	12.91	10.95
						WK16		**19.34**	13.45	16.39	16.81	16.90
UX001-CL301	89[a]	6	2[b]	1.4% (9 / 623)	$P = 0.54$	**WK24**			**27.72**	23.29	18.94	17.14
						WK32				**41.50**	29.96	26.17
						WK40					**34.40**	24.82
						WK48						**32.43**
							WK40	**WK64**				
UX023-CL301 (Pediatrics)	61	2	0	0.5% (1 / 183)	$P = 0.19$	**WK40**	**1.26**	1.14				
						WK64		**1.35**				
							WK4	**WK12**	**WK24**	**WK36**	**WK48**	
						WK4	**594.55**	414.12	72.74	−70.23	−53.46	
						WK12		**591.14**	217.44	68.62	203.40	
UX023-CL304 (Adults)	14	5	1[c]	1.2% (1 / 84)	$P = 0.42$	**WK24**			**239.92**	217.23	336.04	
						WK36				**301.54**	427.17	
						WK48					**659.03**	

MCAR = Missing completely at random; WK = Week

Trials UX001-CL301 and UX023-CL301 were randomized controlled trials, while UX023-CL304 was a single arm trial.

a Of the 89 randomized participants, 88 were included in the prespecified *Primary Analysis Set*, which was defined in the protocol as all randomized participants with a baseline measurement and at least one post-baseline measurement; a single participant did not have any of the 6 post-baseline measurements.

b One participant submitted outcome data for the first five post-BL measurement occasions, while the other participant did not submit any post-BL measurements; reason for study discontinuation was subject non-compliance for both participants.

c The single participant, who withdrew consent, submitted outcome data for the first four post-BL measurement occasions.

d Data are missing completely at random (MCAR) when the pattern of missing values does not depend on the data values. The null hypothesis for Little's MCAR Test is that the data are MCAR. A $P > 0.05$ indicates weak evidence against the null hypothesis.

e The empirical covariance matrix was estimated from the participants included in the prespecified *Primary Analysis Set*, namely, the N = 88, N = 61, and N = 14 participants in trials UX001-CL301, UX023-CL301, and UX023-CL304, respectively, who had a baseline measurement and at least one post-baseline measurement.

The reasons for study discontinuation were subject noncompliance (both participants in UX001-CL301) and withdrawal of consent (single participant in UX023-CL304). In total, 1.4%, 0.5%, and 1.2% of the measurement occasions in UX001-CL301, UX023-CL301, and UX023-CL304, respectively, were missing outcome data. In each trial, no missing outcome data pattern occurred in >1 participant such that the missing outcome data resembled a random sample of all the outcome data. Coupled with the clinical plausibility of MCAR and the results from applying Little's MCAR test (all P>0.19), the longitudinal outcome data within each trial reasonably satisfied the MCAR assumption. Based on the empirical covariances for each outcome, the variances and the pairwise covariances were not constant over time.

Appendix Table 2A shows the operating characteristics associated with each method, with and without the sandwich variance estimator and small-sample correction, based on 100,000 simulated datasets each from UX001-CL301 and UX023-CL301 and 500,000 from UX023-CL304. The prespecified GEE method assumed a CS covariance pattern with the use of the sandwich variance estimator without a small-sample correction and inference based on asymptotic normality, while the data were generated from MMRM with UN covariance pattern. On average, in any given dataset simulated from UX001-CL301, UX023-CL301, and UX023-CL304, 1.4%, 0.5%, and 1.2% of the longitudinal measurements were MCAR.

UX001-CL301

In the largest (N = 88) study with 6 post-baseline measurement occasions and little missing data, incorrect SEs resulted when the sandwich variance estimator was not used with MMRM and a misspecified CS covariance pattern leading to 95% CIs that were too wide or narrow for β_1 and β_{11}, respectively, and power that was too large (relative to the MMRM model with correctly specified covariance) for the test of the primary hypothesis (Figure 1). Specifically, the coverage probability for β_1(between-subject effect) and β_{11} (within-subject effect) was 0.9903 and 0.9274,

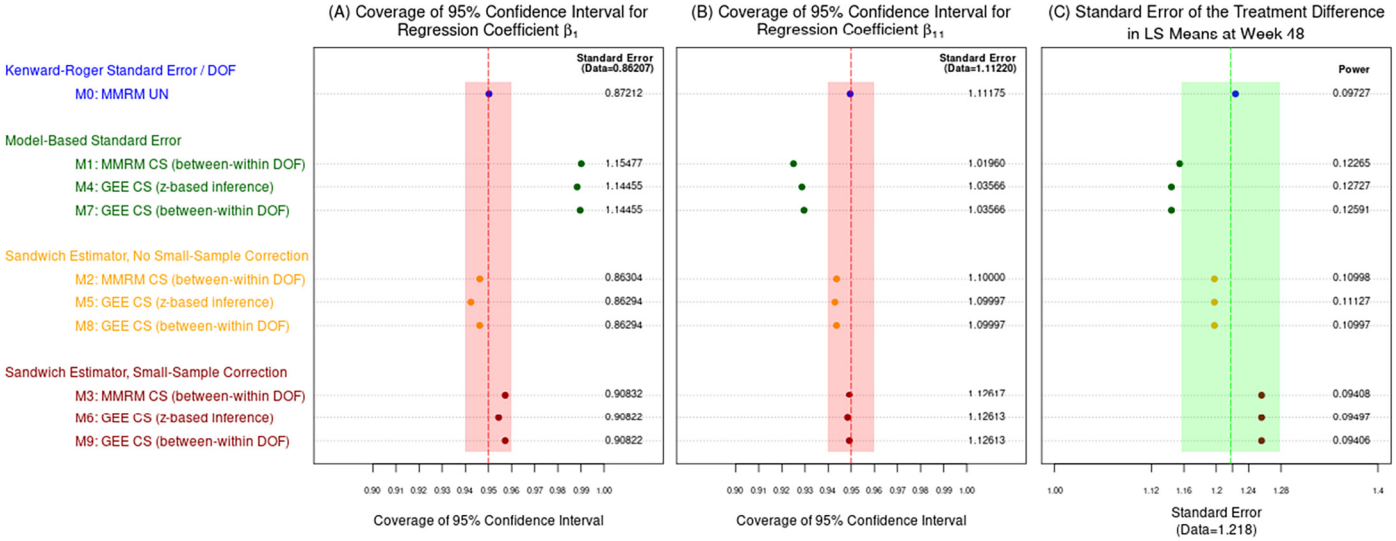


Figure 1: UX001-CL301 (N = 88 Two-Arm Study) – Simulation Results for models M0-M9: Model M0 represents the mixed-model repeated measures (MMRM) model with correctly specified unstructured (UN) covariance and use of the Kenward-Roger method, which combined a standard error adjustment with a Satterthwaite-type degree of freedom (DOF) approximation. M1-M3 represent an MMRM with compound symmetry (CS) covariance pattern that, respectively, used model-based, sandwich-based, and bias-corrected sandwich-based standard errors by Mancl and DeRouen (2001); t-based inference was used with the Between-Within method for determining DOF. M4-M9 represent a generalized estimating equations (GEE) model (marginal model with identify link) with inference based on asymptotic normality (M4-M6) or the t-distribution using the Between-Within DOF-approximation (M7-M9). Plots of the coverage of the 95% confidence intervals for the primary regression coefficients of interest **(A)** β_1 and **(B)** β_{11}, with the average standard errors of the respective regression coefficients listed at the right of each plot; the actual standard error estimated from the trial data for β_1 and β_{11} is shown in parentheses for comparison. The red dashed line on both plots represents the targeted 95% confidence coefficient while the shaded pink region represents coverage between 94% to 96% as a frame of reference. **(C)** Plot of the average standard errors of the treatment difference in the marginal mean change from baseline at week 48 with respect to the actual standard error estimated from the trial data (green dashed line); the shaded green region represents ± 5% of the standard error estimated from the trial data as a frame of reference. The power used to test the primary hypothesis, H_0 : Treatment Difference in LS Means at Week 48 = 0, is listed at the right of the plot.

respectively, and the power testing the primary hypothesis was 0.1238 compared with 0.0977 achieved with MMRM that used a correctly specified covariance pattern and the Kenward-Roger adjustment. Findings were similar when the model-based SEs were used with GEE. Although the bias-corrected SEs for β_1, β_{11}, and for the primary hypothesis test with GEE and MMRM were larger, relative to the SEs seen in the data, t-based inference yielded coverage probabilities $\geq$95% for both β_1 and β_{11}, and power that was commensurate with the power achieved with MMRM that used a correctly specified covariance pattern and the Kenward-Roger adjustment.

UX023-CL301

In the modestly large study (N = 61) with 2 post-baseline measurement occasions and little missing data, any differences between the methods were subtle but no less instructive (Figure 2). A CS covariance pattern was a close approximation to the true underlying covariance pattern that was used to simulate the data. The elements of the 2x2 covariance matrix used in generating the data were $\sigma_1^2 = 0.5992$, $\sigma_2^2 = 0.5795$, and $\sigma_{12} = \sigma_{12} = 0.4437$. Consequently, both methods with model-based SEs and t-based inference achieved 95% coverage for β_1 (between-subject effect). MMRM with the Kenward-Roger adjustment also achieved the correct coverage. The use of the sandwich variance estimator without and with the bias-correction resulted in SEs that were modestly smaller and larger, respectively, relative to the SEs of β_1 and in the test of the primary hypothesis that was seen in the data.

UX023-CL304

In the small, single arm study (N = 14) with 5 post-baseline measurement occasions and for which there were also few missing data points, the CS covariance pattern was markedly misspecified for all models studied. Using bias-

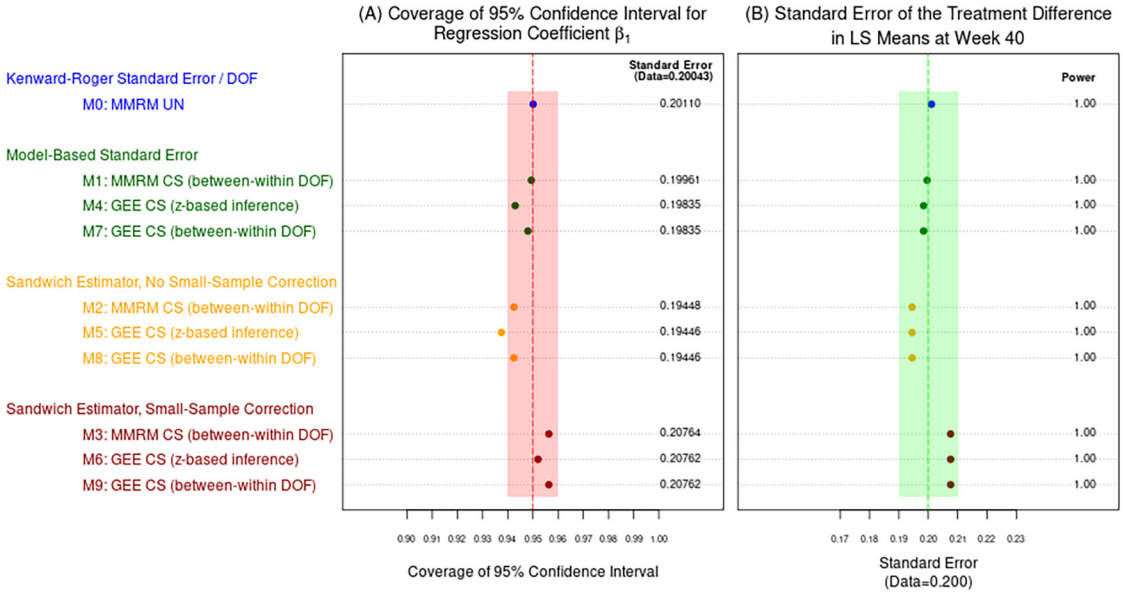


Figure 2: UX023-CL301 (N = 61 Two-Arm Study) – Simulation Results for models M0-M9: Model M0 represents the mixed-model repeated measures (MMRM) model with correctly specified unstructured (UN) covariance and use of the Kenward-Roger method, which combined a standard error adjustment with a Satterthwaite-type degree of freedom (DOF) approximation. M1-M3 represent an MMRM with compound symmetry (CS) covariance pattern that, respectively, used model-based, sandwich-based, and bias-corrected sandwich-based standard errors by Mancl and DeRouen (2001); t-based inference was used with the Between-Within method for determining DOF. M4-M9 represent a generalized estimating equations (GEE) model (marginal model with identify link) with inference based on asymptotic normality (M4-M6) or the t-distribution using the Between-Within DOF-approximation (M7-M9). Plots of the coverage of the 95% confidence intervals for the primary regression coefficient of interest **(A)** β_1, with the average standard errors of the respective regression coefficient listed at the right of the plot; the actual standard error estimated from the trial data for β_1 is shown in parentheses for comparison. The red dashed line on the plot represents the targeted 95% confidence coefficient while the shaded pink region represents coverage between 94% to 96% as a frame of reference. **(B)** Plot of the average standard errors of the treatment difference in the marginal mean change from baseline at week 40 with respect to the actual standard error estimated from the trial data (green dashed line); the shaded green region represents $\pm$ 5% of the actual standard error estimated from the trial data as a frame of reference. The power used to test the primary hypothesis, H_0 : Treatment Difference in LS Means at Week 40 = 0, is listed at the right of the plot.

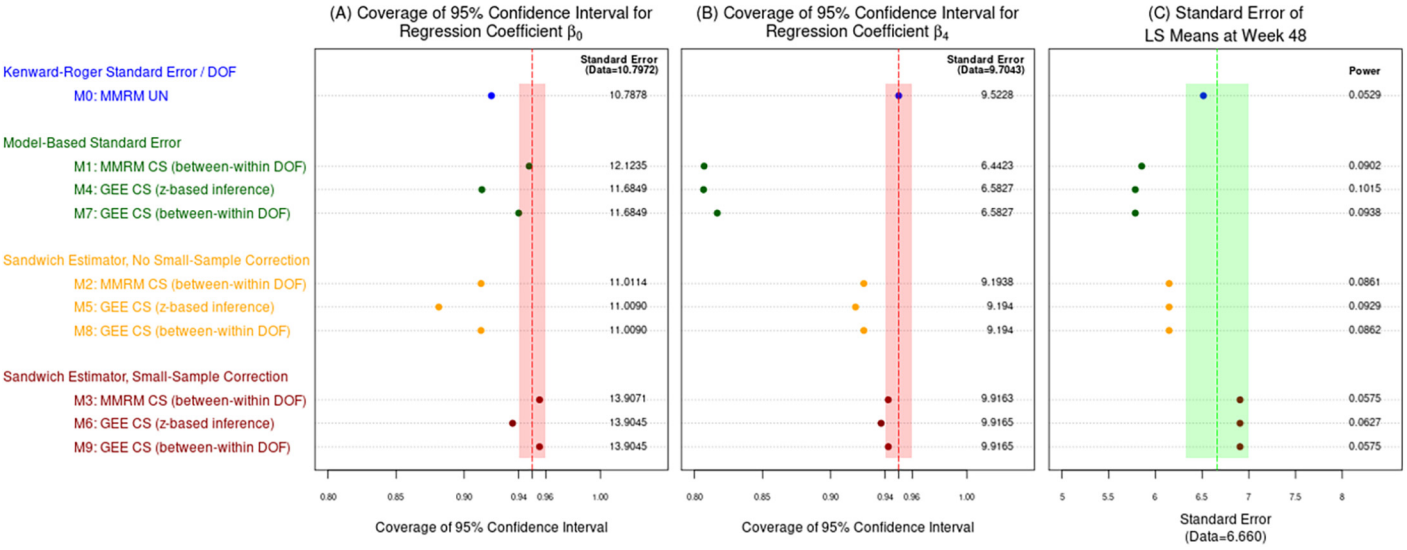


Figure 3: UX023-CL304 (N = 14 Single-Arm Study) – Simulation Results for models M0-M9: Model M0 represents the mixed-model repeated measures (MMRM) model with correctly specified unstructured (UN) covariance and use of the Kenward-Roger method, which combined a standard error adjustment with a Satterthwaite-type degree of freedom (DOF) approximation. M1-M3 represent an MMRM with compound symmetry (CS) covariance pattern that, respectively, used model-based, sandwich-based, and bias-corrected sandwich-based standard errors by Mancl and DeRouen (2001); t-based inference was used with the Between-Within method for determining DOF. M4-M9 represent a generalized estimating equations (GEE) model (marginal model with identify link) with inference based on asymptotic normality (M4-M6) or the t-distribution using the Between-Within DOF-approximation (M7-M9). Plots of the coverage of the 95% confidence intervals for the primary regression coefficients of interest **(A)** β_0, and **(B)** β_4 with the average standard errors of the respective regression coefficients listed at the right of each plot; the actual standard error estimated from the trial data for each regression coefficient is shown in parentheses for comparison. The red dashed line represents the targeted 95% confidence coefficient while the shaded pink region represents coverage between 94% to 96% as a frame of reference. **(C)** Plot of the average standard errors of the marginal mean change from baseline at week 48 with respect to the actual standard error estimated from the trial data (green dashed line); the shaded green region represents ± 5% of the actual standard error estimated from the trial data as a frame of reference. The power used to test the primary hypothesis, $H_0: LS$ Means at Week 48 = 0, is listed at the right of the plot.

corrected sandwich variance estimators with both methods and with t-based inference was needed, consistently achieving coverage probabilities closest to 95% (β_0 [between-subject effect] and β_4[within-subject effect]: 0.9541 and 0.9417 with both MMRM and GEE using t-based inference). Additionally, the standard error and power for the test of the primary hypothesis was similar to that achieved by MMRM with correctly specified covariance and the Kenward-Roger adjustment. Other models that assumed a CS covariance pattern consistently performed considerably worse (Figure 3). Interestingly, for MMRM with correctly specified covariance, the Kenward-Roger adjustment to the standard error for the intercept, β_0, was consistent with the data, however, the Satterthwaite-type DOF approximation based on that adjusted standard error resulted in 92% coverage. For all other regression coefficients (β_1, β_2, β_3, and β_4), which were fixed effects that changed within subjects, the correct coverage was achieved (Appendix Table 2).

3.1 Summary of Results

In each trial, <1.5% of the measurement occasions were missing outcome data and the MCAR assumption was reasonably satisfied. Box 1 succinctly summarizes the main findings. Notably, the empirical variances and pairwise covariances were not constant over time in UX001-CL301 and UX023-CL304; assuming a CS covariance pattern, therefore, resulted in considerable model misspecification of the covariances. Consequently, MMRM and GEE assuming a CS covariance pattern without the use of the sandwich variance estimator performed badly. Bias-corrected SEs with t-based inference was needed in the smallest trial, UX023-CL304, with both methods to consistently achieve at least 94% coverage probability for the regression coefficients; furthermore, the power testing the primary hypothesis was commensurate with the power achieved with MMRM that used a correctly specified covariance pattern and the Kenward-Roger standard error adjustment. The largest trial, UX001-CL301,

Box 1. Trial-specific simulation summary highlights when using a compound symmetry (CS) covariance pattern.

UX001-CL301 (N = 88 Two-Arm Study; 6 post-baseline occasions)

- In the presence of model misspecification of the covariances, model-based standard errors should not be used even with MMRM. Sandwich variance estimators were needed for both the MMRM and GEE methods, while a bias-correction to those sandwich variance estimators conservatively achieved ≥95% coverage probability and power testing the primary hypothesis, H_0 : Treatment Difference in *LS* Means at Week 48 = 0, that was commensurate with the power achieved with MMRM that used a correctly specified covariance pattern and the Kenward-Roger standard error adjustment.

UX023-CL301 (N = 61 Two-Arm Study; 2 post-baseline occasions)

- Given that the compound symmetry covariance pattern was a close approximation to the truth, model-based SEs with *t*-based inference with the MMRM or GEE methods or with the MMRM with the Kenward-Roger standard error adjustment achieved the correct coverage probabilities. If the sandwich variance estimator is used, then a bias-correction is recommended to achieve ≥95% coverage probability for the regression coefficients.

UX023-CL304 (N = 14 Single-Arm Study; 5 post-baseline occasions)

- In the presence of model misspecification of the covariances, the sandwich variance estimator with the bias-correction and with *t*-based inference was needed for both the MMRM and GEE methods to consistently achieve at least 94% coverage probability for the regression coefficients; furthermore, the power testing the primary hypothesis, H_0 : *LS* Means at Week 48 = 0, was commensurate with the power achieved with MMRM that used a correctly specified covariance pattern and the Kenward-Roger standard error adjustment.

CONCLUSION

In each case study, the missing completely at random (MCAR) assumption was plausible and the missingness low. When modeling the mean response with time as a categorical variable, MMRM with unstructured (UN) covariance coupled with the Kenward-Roger standard error adjustment should be used if the data allow it. If not, then a parsimonious CS covariance structure can be considered. In the two case studies that exhibited nonconstant variance/covariances over time (**UX001-CL301** and **UX023-CL304**), the sandwich variance estimator and small-sample correction with *t*-based inference using the between-within degree of freedom (DOF) method should be considered with both MMRM and GEE. If the CS pattern was a good approximation, as seen in **UX023-CL301**, then model-based standard errors with *t*-based inference using the between-within DOF performed well for both methods.

also benefited from bias-corrected SEs with t-based inference that yielded coverage probabilities ≥95% and power identical to the power achieved when the covariance pattern was correctly specified. In the modestly large study, UX023-CL301, with only two post-baseline measurement occasions, adopting the sandwich variance estimator was not helpful. Given that the CS covariance pattern was a close approximation to the truth, the larger sampling variability associated with the sandwich variance estimator [17] resulted in coverage probabilities that were less than the nominal level; choosing model-based SEs with *t*-based inference with either method or MMRM with the Kenward-Roger adjustment achieved the correct coverage probability.

4. DISCUSSION

MMRM, a response profile analysis, arguably dates back to Greenhouse and Geisser (1959) [38], and has been widely adopted as the primary analysis for analyzing repeatedly measured continuous outcomes in randomized clinical trials (RCTs). For an example of a recently completed phase 3 trial (N = 65; NCT03399786) in a rare disease that applied MMRM as the primary analysis to assess a treatment effect at a single time point, albeit without the sandwich variance estimator and small-sample correction, see Raal et al (2020) [39]. Since the two initial companion papers on GEE for longitudinal data [3, 40] coupled with the connections drawn between GEE and likelihood-based methods [41] – and the widely available software notwithstanding – GEE has infrequently been used as the primary analysis for continuous outcomes in RCTs. Based on a recent comprehensive evaluation of statistical methods applied as the primary analysis to repeatedly measured continuous outcomes from RCTs published in 2019, Ren et al (2022) [42] estimated 4% (4/96) adopted GEE to assess a treatment effect at a single time point compared with 36% (35/96) that used MMRM; the remaining 59% (57/96) oversimplified the longitudinal design and applied conventional methods such as t-tests and analysis of covariance to assess a treatment effect at a single time point. One rare disease trial (N = 53; NCT02208687) that adopted in primary analysis GEE with the sandwich variance estimator was designed to investigate the effectiveness at a single time point of a self-management group program

to improve social participation and endurance in patients with neuromuscular disease with chronic fatigue [43]. Either method can be considered for modeling longitudinal changes in the mean response in trials conducted in rare diseases with a parsimonious CS covariance structure; however, care needs to be exercised with both methods to ensure its correct use.

Inferences about longitudinal changes in mean response and its relationship to treatment arm are sensitive to the chosen covariance model. This is true for both MMRM and GEE when modeling the mean with time included categorically, and despite the covariances being treated as nuisance parameters with GEE. In rare disease trials, it is common, however, to pragmatically prespecify a parsimonious CS covariance pattern due to the smaller sample sizes or to prespecify a decision to adopt a CS covariance pattern if an UN covariance pattern was inestimable. Based on our case studies, the variances and covariances of the longitudinal outcomes were not constant over time. When the CS covariance pattern does not closely approximate the true but unknown covariance pattern, incorrect SEs of the regression coefficients will result. As seen with our case studies, to obtain valid SEs, the sandwich variance estimator coupled with a small-sample correction to correct for bias, must be considered with both methods.

With GEE, the sandwich variance estimator – as opposed to model-based SEs – is the default setting in most statistical packages, however, the opposite is true when fitting MMRM. Consider, for example, SAS and R. In SAS, the user needs to specify the EMPIRICAL option in the PROC GLIMMIX (or MIXED) statement to apply the sandwich variance estimator when fitting MMRM, while the sandwich variance estimator is the default when using GENMOD to fit the corresponding GEE. When fitting MMRM in R using the *gls* function in the *nlme* package [44], the analyst would inconveniently appeal to the *clubSandwich* package [45] and the function *vcovCR*, which returns a sandwich estimate of the variance-covariance matrix of the regression coefficients. The recently developed *mmrm* package [13] currently does not have an option to apply sandwich variance estimators. To fit the corresponding GEE in R, the user can appeal to the *gee* [46] or *geepack* [47] packages, and in both packages the sandwich variance estimator is the default. Defaulting to the sandwich variance estimator with GEE is not surprising. The properties of the sandwich variance estimator has been well studied with GEE [19, 48–51]. If there is misspecification of the covariance, however, not using the sandwich variance estimator can also result in misleading inferences for the regression coefficients in MMRM as seen in our case studies.

Some small-sample corrections to the sandwich variance estimators are available in SAS and R to correct for bias along with some DOF approximations to the t distribution. Pustejovsky and Tipton (2017) [24] and Wang and coauthors (2016) [25] provide a comprehensive review of small-sample corrections available in R and guidance on which approach to use for the methods considered in this article. Gosho et al (2021) [22] provide a practical review of the small-sample corrections for MMRM that can be obtained using the GLIMMIX procedure directly or programmatically using IML.

Longitudinal responses will invariably be incomplete. Unlike MCAR, when the missing response data mechanism is MAR, subjects with missing response data will no longer be a random subset of the sample. It is well known that MMRM provides a valid framework under MAR, a more defensible assumption than MCAR in any given trial, albeit provided that the covariance pattern has been correctly specified [1]. If estimating an UN covariance pattern is prohibitive due to the small sample, resulting in misspecification, inferences about the mean response over time will be invalid when dropout is MAR. This point further underscores the importance of considering the sandwich variance estimator with MMRM ($\pm$ a small-sample correction). When dropout is MAR as opposed to MCAR, the validity of the analysis with the standard GEE method based on available observations will be compromised [14]. Inverse probability weighted (IPW) [10] GEE should be adopted to adjust the analysis for the propensity to dropout; otherwise, GEE will yield biased estimates of the coefficients in the model part of the mean response. The sandwich variance estimator when using IPW methods can be constructed to adjust for estimation of the weights. Because most statistical procedures allow for the inclusion of sampling weights, it is straightforward to apply IPW-GEE. While unlikely in a well-controlled longitudinal clinical trial [1], if the longitudinal responses are missing not at random (MNAR), none of the methods considered in this article will be valid [11].

In our case studies, the MCAR assumption was clinically defensible. The number of measurement occasions including baseline was $\leq$7, few patients discontinued prematurely, no missing outcome data pattern occurred in >1 participant, and the degree of missing outcome data was <1.5%. While the statistical test of the MCAR assumption likely had low power [52], its use combined with the empirical evidence provided justification to use GEE. Due to the hierarchy of dropout mechanisms [53], MMRM was also justified.

There were research limitations. First, we only considered GLIMMIX to estimate MMRM and the SAS/IML GEE macro, which reproduces the method-of-moments approach to estimation with GENMOD. SAS is widely used software and the primary analysis in each trial used GENMOD. Second, there are numerous bias-corrections to sandwich variance estimators in the literature but only a few are currently available with GLIMMIX and none with GENMOD. The choice of which bias-correction to use would depend on several factors [19, 24, 25] and should be prespecified. In this article, we only considered the bias-corrected SEs recommended by Mancl and DeRouen (2001) [19]. Third, for t-based inference, we only considered the between-within method DOF approximation for small-sample inference recommended by Schluchter and Elashoff (1990) [36]. When

empirical or bias-corrected SEs are used with GLIMMIX, DOF approximation methods are limited to between-within, containment, and residual, where the latter two are not small-sample approximations; furthermore, GENMOD only performs inference assuming asymptotic normality. The determination of which DOF approximation to use is an active area of research, with many methods not accessible in standard SAS and R routines. Lastly, the research was limited to three case studies and, in each, <1.5% of the measurement occasions were missing outcome data. Although these case studies represented typical longitudinal designs in rare diseases, it was not an exhaustive representation and the conclusions may not necessarily extend to other design variations. In particular, dropout other than MCAR was not considered in the simulation study. However, to understand whether the findings held in the presence of moderate missingness defined as 10% of the measurement occasions being MCAR, the simulation study was repeated. The simulation-based conclusions when using a CS covariance pattern remained; however, for the smallest trial (UX023-CL304; N = 14), estimation was unstable or not always possible when assuming an unstructured covariance pattern. In any case, the intent of the current research was not to pressure test these methods; rather to elucidate the issues and the steps to address the issues using real datasets in rare diseases to aid an analyst working in a rare disease in selecting an appropriate modeling strategy.

5. CONCLUSION

Based on a review of three case studies recently conducted in rare diseases, the MCAR assumption was plausible and the missingness low. When modeling the mean response with time as a categorical variable, MMRM with UN covariance coupled with the Kenward-Roger standard error adjustment should be used if the data allow it. If not, then a parsimonious CS covariance structure can be considered. In the two case studies that exhibited nonconstant variance/covariances over time, the sandwich variance estimator and small-sample correction with t-based inference using the between-within DOF should be considered with both MMRM and GEE. If the CS pattern was a good approximation, as seen in the other case study, then model-based standard errors with t-based inference using the between-within DOF performed well with both methods.

DECLARATION OF ANY POTENTIAL CONFLICTS OF INTEREST

At the time the research was conducted, David Zahrieh, Yi Wang, and Tony Koutsoukos were employees at Ultragenyx Pharmaceutical Inc. This research reflects the views of the authors and should not be construed to represent Ultragenyx's views or policies.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this research.

PRESENTATIONS

This work was presented as a Poster at the 2023 Joint Statistical Meetings on August 7, 2023.

Accepted 8 December 2025

REFERENCES

[1] Mallinckrodt, C. H., Lane, P. W., Schnell, D., Peng, Y. and Mancuso, J. P. Recommendations for the primary analysis of continuous endpoints in longitudinal clinical trials. *Drug Inf J.* **42**(4) 303–319 (2008). https://doi.org/10.1177/009286150804200402

[2] Mancl, L. A. and Leroux, B. G. Efficience of regression estimates for clustered data. *Biometrics.* **52** 500–511 (1996).

[3] Liang, K. Y. and Longitudinal, Z. SL. Data-Analysis Using Generalized Linear-Models. *Biometrika.* **73**(1) 13–22 (1986). https://doi.org/10.1093/biomet/73.1.13. MR0836430

[4] Day, S., Jonker, A. H., Lau, L. P. L. et al. Recommendations for the design of small population clinical trials. *Orphanet J Rare Dis.* **13**(1) 195 (2018). https://doi.org/10.1186/s13023-018-0931-2

[5] CMW, v. d. W. M. and du Prie-Olthof MJ, G. The patient's view on rare disease trial design - a qualitative study. *Orphanet J Rare Dis.* **14**(1) 31 (2019). https://doi.org/10.1186/s13023-019-1002-z

[6] Hall, A. K. and Ludington, E. Considerations for successful clinical development for orphan indications. *Expert Opinion on Orphan Drugs.* **1**(11) 847–850 (2013).

[7] Hilgers, R. D., König, F., Molenberghs, G. and Senn, S. J. Design and analysis of clinical trials for small rare disease populations. *J Rare Dis Res Treat.* **1**(3) 53–60 (2016).

[8] Abrahamyan, L., Feldman, B. M., Tomlinson, G. et al. Alternative designs for clinical trials in rare diseases. *Am J Med Genet C Semin Med Genet.* **172**(4) 313–331 (2016). https://doi.org/10.1002/ajmg.c.31533

[9] Pizzamiglio, C., Vernon, H. J., Hanna, M. G. and Pitceathly, R. DS. Designing clinical trials for rare diseases: unique challenges and opportunities. *Nat Rev Methods Primers.* **2**(1) (2022). https://doi.org/10.1038/s43586--022-00100-2

[10] Robins, J. M., Rotnitzky, A. and Zhao, L. P. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association.* **90** 106–121 (1995). MR1325118

[11] Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. *Applied longitudinal data analysis.* 2nd ed. *Probability and statistics.* John Wiley & Sons (2011). MR2830137

[12] Gurka, M. J. Selecting the best linear mixed model under REML. *The American Statistician.* **60** 19–26 (2006). https://doi.org/10.1198/000313006X90396. MR2224133

[13] mmrm: Mixed models for repeated measures (2022). https://CRAN.R-project.org/package=mmrm

[14] Lu, K. and Mehrotra, D. V. Specification of covariance structure in longitudinal data analysis for randomized clinical trials. *Stat Med.* **29**(4) 474–488 (2010). https://doi.org/10.1002/sim.3820. MR2751783

[15] Liang, K. Y. and Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika.* **73**(1) 13–22 (1986). https://doi.org/10.1093/biomet/73.1.13. MR0836430

[16] Gurka, M. J., Edwards, L. J. and Muller, K. E. Avoiding bias in mixed model inference for fixed effects. *Stat Med.* **30**(22) 2696–2707 (2011). https://doi.org/10.1002/sim.4293. MR2843173

[17] Kauermann, G. and Carroll, R. J. The sandwich variance estimator: Efficiency properties and coverage probability of confidence intervals. Discussion Paper 189 Collaborative Research Center 3862000.
[18] Kauermann, G. and Carroll, R. J. A note on the efficiency of sandwich covariance matrix estimation. *J Am Stat Assoc.* **96**(456) 1387–1396 (2001). https://doi.org/10.1198/016214501753382309. MR1946584
[19] Mancl, L. A. and DeRouen, T. A. A covariance estimator for GEE with improved small-sample properties. *Biometrics.* **57**(1) 126–134 (2001). https://doi.org/10.1111/j.0006-341x.2001.00126.x. MR1833298
[20] Bell, R. M. and McCaffrey, D. F. Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology.* **28** 169–181 (2002).
[21] Gosho, M., Hirakawa, A., Noma, H., Maruo, K. and Sato, Y. Comparison of bias-corrected covariance estimators for MMRM analysis in longitudinal data with dropouts. *Stat Methods Med Res.* **26**(5) 2389–2406 (2017). https://doi.org/10.1177/0962280215597938. MR3712239
[22] Gosho, M., Noma, H. and Maruo, K. Practical Review and Comparison of Modified Covariance Estimators for Linear Mixed Models in Small-sample Longitudinal Studies with Missing Data. *Int Stat Rev.* **89**(3) 550–572 (2021). https://doi.org/10.1111/insr.12447. MR4411918
[23] Gosho, M., Sato, T. and Takeuchi, H. Robust covariance estimator for small-sample adjustment in the generalized estimating equations: a simulation study. *Science Journal of Applied Mathematics and Statistics.* **2**(1) 20–25 (2014).
[24] Pustejovsky, J. E. and Tipton, E. Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business and Economic Statistics.* **36**(4) 672–683 (2018). https://doi.org/10.1080/07350015.2016.1247004. MR3871709
[25] Wang, M., Kong, L., Li, Z. and Zhang, L. Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples. *Stat Med.* **35**(10) 1706–1721 (2016). https://doi.org/10.1002/sim.6817. MR3513479
[26] Imbens, G. W. and Kolesar, M. Robust standard errors in small samples: Some practical advice. In *Research NBoE, editor. NBER Working Paper Series. National Bureau of Economic Research* (2012). Cambridge, MA.
[27] Pan, W. On the robust variance estimator in generalised estimating equations. *Biometrika.* **88**(3) 901–906 (2001). https://doi.org/10.1093/biomet/88.3.901. MR1859421
[28] ICH E9 Statistical Principles for Clinical Trials (1998).
[29] Lochmüller, H., Behin, A., Caraco, Y. et al. A phase 3 randomized study evaluating sialic acid extended-release for GNE myopathy. *Neurology.* **92**(18) E2109–E2117 (2019). https://doi.org/10.1212/Wnl.0000000000006932
[30] Imel, E. A., Glorieux, F. H., Whyte, M. P. et al. Burosumab versus conventional therapy in children with X-linked hypophosphataemia: a randomised, active-controlled, open-label, phase 3 trial. *Lancet.* **393**(10189) 2416–2427 (2019). https://doi.org/10.1016/S0140-6736(19)30654-3
[31] Insogna, K. L., Rauch, F., Kamenicky, P. et al. Burosumab Improved Histomorphometric Measures of Osteomalacia in Adults with X-Linked Hypophosphatemia: A Phase 3, Single-Arm. *International Trial. J Bone Miner Res.* **34**(12) 2183–2191 (2019). https://doi.org/10.1002/jbmr.3843
[32] Little, R. J. A. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association.* **83**(44) 1198–1202 (1988). MR0997603
[33] Akaike, H. In *Information theory and an extension of the maximum likelihood principle* 267–281 (1973). MR0483125
[34] Burnham, K. P. and Anderson, D. R. *Model selection and multimodel inference - A practical information-theoretic approach.* 2 ed. Springer, New York, NY (2002). MR1919620
[35] Kenward, M. G. and Roger, J. H. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics.* **53**(3) 983–997 (1997). https://doi.org/10.2307/2533558
[36] Schluchter, M. D. and Small-Sample, E. JD. Adjustments to Tests with Unbalanced Re peated Measures Assuming Several Covariance Structures. *Journal of Statistical Computation and Simulation.* **37** 69–87 (1990).
[37] Hammill, B. G. and Preisser, J. S. A SAS/IML software program for GEE and regression diagnostics. *Comput Stat Data An.* **51**(2) 1197–1212 (2006). https://doi.org/10.1016/j.csda.2005.11.016. MR2297517
[38] Greenhouse, S. W. and Geisser, S. On methods in the analysis of profile data. *Psychometrika.* **32** 95–112 (1959). https://doi.org/10.1007/BF02289823. MR0103783
[39] Raal, F. J., Rosenson, R. S., Reeskamp, L. F. et al. Evinacumab for Homozygous Familial Hypercholesterolemia. *N Engl J Med.* **383**(8) 711–720 (2020). https://doi.org/10.1056/NEJMoa2004215
[40] Zeger, S. L. and Liang, K. Y. Longitudinal Data-Analysis for Discrete and Continuous Outcomes. *Biometrics.* **42**(1) 121–130 (1986). https://doi.org/10.2307/2531248
[41] Zhao, L. P., Prentice, R. and Self, S. Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society, Series B.* **54** 805–811 (1992).
[42] Ren Zhang Y Jia Y, Y. et al. Analyses of repeatedly measured continuous outcomes in randomized controlled trials needed substantial improvements. *J Clin Epidemiol.* **143** 105–117 (2022). https://doi.org/10.1016/j.jclinepi.2021.12.007
[43] Veenhuizen, Y., Cup, E. H. C., Jonker, M. A. et al. Self-management program improves participation in patients with neuromuscular disease: A randomized controlled trial. *Neurology.* **93**(18) e1720–e1731 (2019). https://doi.org/10.1212/WNL.0000000000008393
[44] nlme: Linear and Nonlinear Mixed Effects Models. Version Version R package version 3.1-163 (2023). https://CRAN.R-project.org/package=nlme
[45] clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections. Version R package version 0.5.8.9999 (2023). http://jepusto.github.io/clubSandwich/
[46] gee: Generalized estimation equation solver. Version R package version 4.13-19 (2015). https://cran.r-project.org/web/packages/gee/
[47] Halekoh, U., Hojsgaard, S. and Yan, J. The R Package geepack for Generalized Estimating Equations. *J Stat Softw.* **15**(2) 1–11 (2006). https://doi.org/10.18637/jss.v015.i02
[48] Fay, M. P. and Graubard, B. I. Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics.* **57**(4) 1198–1206 (2001). https://doi.org/10.1111/j.0006-341X.2001.01198.x. MR1950428
[49] Gunsolley, J. C., Getchell, C. and Chinchilli, V. M. Small sample characteristics of generalized estimating equations. *Communications in Statistics, Simulation and Computation.* **24** 869–878 (1995).
[50] Hinkley, D. V. and Wang, S. Efficiency of robust standard errors for regression coefficients. *Communications in Statistics, Theory and Methods.* **20** 1–11 (1991). https://doi.org/10.1080/03610929108830479. MR1114631
[51] Kauermann, G. and Carroll, R. J. The sandwich variance estimator: efficiency properties and coverage probability of confidence intervals. *Journal of the American Statistical Association.* **96** 1387–1396 (2001). https://doi.org/10.1198/016214501753382309. MR1946584
[52] Thoemmes, F. and Enders, C. K. A structural equation model for testing whether data are missing completely at random presented at. *Annual Meeting of the American Educational Research Association, Chicago* Chicago, IL (2007).
[53] Fitzmaurice, G. M. Methods for handling dropouts in longitudinal clinical trials. *Statistica Neerlandica.* **57**(1) 75–99 (2003). https://doi.org/10.1111/1467-9574.00222. MR2055522

Appendix Table 1. True values for the regression coefficients and variance/covariance matrix used to generate the data in the simulation study.

UX001-CL301			UX023-CL304 (Adults)			UX023-CL301 (Pediatrics)		
Parameter	**Variable**	**MMRM UN Estimate (SE)**	**Parameter**	**Variable**	**MMRM UN Estimate (SE)**	**Parameter**	**Variable**	**MMRM UN Estimate (SE)**
β_0	Intercept	-0.8352952 (1.1847466)	β_0	Intercept	24.21299 (10.7972)	β_0	Intercept	-0.31830536 (0.3458454)
β_1	TRT01P6 g/day SA-ER	0.8337131 (0.8620675)	β_1	AVISITN12	-3.07143 (5.0530)	β_1	TRT01AN Conv. Therapy	1.35857205 (0.2004256)
β_2	AVISITN16	0.4537376 (0.6615980)	β_2	AVISITN24	-17.28571 (7.0152)	β_2	AVISITN64	-0.15212510 (0.1018131)
β_3	AVISITN24	-0.1595676 (0.7639032)	β_3	AVISITN36	-20.35714 (8.6046)	β_3	TRT01AN Conv. Ther.:AVISITN64	-0.14474990 (0.1395341)
β_4	AVISITN32	-1.3867574 (0.8081185)	β_4	AVISITN48	-20.17305 (9.7043)	γ_1	TOTSCORE	-0.55749136 (0.0890778)
β_5	AVISITN40	-1.3269535 (0.7662077)	γ_1	BASE	-0.02524 (0.0749)	γ_2	factor(AGEGRPN) ≥ 5	0.02877922 (0.1900769)
β_6	AVISITN48	-2.4196390 (0.7954184)						
β_7	TRT01P6 g/day SA-ER:AVISITN16	-1.5581821 (0.9219093)						
β_8	TRT01P6 g/day SA-ER:AVISITN24	-0.5470991 (1.0626885)						
β_9	TRT01P6 g/day SA-ER:AVISITN32	0.8012018 (1.1275762)						
β_{10}	TRT01P6 g/day SA-ER:AVISITN40	0.3280646 (1.0686359)						
β_{11}	TRT01P6 g/day SA-ER:AVISITN48	-0.0377098 (1.1122039)						
γ_1	BASE	0.0060925 (0.0139584)						
γ_2	factor(SEX)M	0.0441477 (0.7748458)						
γ_3	factor(REGION)The Rest of the World	-0.5190236 (0.7978921)						
Variance/Covariance Matrix, Σ (Unstructured) **16.1450** 8.6339 9.820 14.505 12.556 10.830 8.6339 **19.6720** 13.782 16.616 17.133 17.188 9.8200 13.7820 **28.054** 23.133 18.717 17.065 14.5050 16.6160 23.133 **40.691** 29.298 25.589 12.5560 17.1330 18.717 29.298 **33.937** 24.264 10.8300 17.1880 17.065 25.589 24.264 **32.438**			**Variance/Covariance Matrix, Σ (Unstructured)** **626.790** 447.38 92.78 -54.604 -35.735 447.380 **625.42** 238.50 85.260 202.580 92.780 238.50 **247.76** 220.660 313.100 -54.604 85.26 220.66 **300.560** 393.610 -35.735 202.58 313.10 393.610 **613.850**			**Variance/Covariance Matrix, Σ (Unstructured)** **0.59917** 0.44368 0.44368 **0.57952**		
Covariates The covariates BASE, SEX (Male vs Female), and REGION (The Rest of the World vs US) were generated from a distribution reasonably consistent with the trial data, namely, a truncated *N*(56.14, 27.954) for BASE, 55% males, and 66% rest of the world.			**Covariates** The covariates BASE was generated from a distribution reasonably consistent with the trial data, namely, a truncated *N*(113.14, 51.600).			**Covariates** The covariates TOTSCORE and AGEGRPN (≥ 5 vs <5) were generated from a distribution reasonably consistent with the trial data, namely, a truncated *N*(3.18, 1.057) for TOTSCORE and 57% ≥ 5 years for AGEGRPN.		
Difference (Placebo - Treatment) in LS Means at Week 48 (SE) -0.796 (1.218)			**LS Mean at Week 48 (SE)** 1.21 (6.66)			**Difference (Burosumab – Conv. Therapy) in LS Means at Week 40** -1.36 (0.200)		

Note. The outcomes modeled in UX001-CL301 (N=88), UX023-CL304 (N=14), and UX023-CL301 (N=61) were the changes from baseline in the HDD upper extremity composite score, changes from baseline in alkaline phosphatase, and changes from baseline in rickets severity. The standard errors of the regression coefficients were calculated by taking the square root of the diagonal elements from the covariance matrix of the regression coefficients, which was calculated as Cov $(\beta)= [\sum_{i=1}^{N}(X_i^T\Sigma_i^{-1}X_i)]^{-1}$; here, X denotes the design matrix and Σ the variance/covariance matrix for the i^{th} subject. For the MMRM model applied to each study dataset, the AIC was used to gauge whether the criterion was expressing a preference for the unstructured or compound symmetry covariance structure. With the trial datasets for UX001-CL301 and UX023-CL304, the AIC expressed a strong preference for the unstructured covariance pattern (difference > 10), while with the trial dataset for UX023-CL301 the AIC expressed a preference for the compound symmetry covariance pattern (difference < 2).

Appendix Table 2A. Simulation study and comprehensive results for study UX001-CL301.

Study UX001-CL301
N = 88, Two Arm Study

Data Generating Model

MMRM Model: $E(Y_i|X_i) = \beta_0 + \beta_1 I(\mathrm{G}=1) + \beta_2 I(\mathrm{W}=16) + \beta_3 I(\mathrm{W}=24) + \beta_4 I(\mathrm{W}=32) + \beta_5 I(\mathrm{W}=40) + \beta_6 I(\mathrm{W}=48) + \beta_7 I(\mathrm{G}=1)\times I(\mathrm{W}=16)$
$+ \beta_8 I(\mathrm{G}=1)\times I(\mathrm{W}=24) + \beta_9 I(\mathrm{G}=1)\times I(\mathrm{W}=32) + \beta_{10} I(\mathrm{G}=1)\times I(\mathrm{W}=40) + \beta_{11} I(\mathrm{G}=1)\times I(\mathrm{W}=48)$
$+\gamma_1 \mathrm{Base} + \gamma_2 I(\mathrm{Sex}=\mathrm{Male}) + \gamma_3 I(\mathrm{Region}=\mathrm{NonUS})$

Repeated measurements taken on $i = 1, \ldots, 88$ participants (45 treatment; 43 placebo) at the same set of 6 post-baseline occasions.
Y_i = Change from baseline.
I represents an indicator variable.
G = Group; W = Week; Base = Baseline value; Reference categories for variables G, W, Sex, and Region were placebo, week 8, female, and US, respectively.
Unstructured (UN) covariance structure was assumed.
The estimates of the regression coefficients and variance / covariance matrix from fitting the MMRM model to the trial data were used as the true values.
The covariates Base, Sex, and Region were generated from a distribution reasonably consistent with the trial data: a truncated *N*(56.14, 27.954), 55% males, and 66% Non-US.
100,000 data sets generated; on average, in any given simulated data set, 1.4% of the data were missing completely at random.
When the Kenward-Roger adjusted standard errors were used, Satterthwaite-type degrees of freedom (DOF) were computed based on this adjustment and used to construct approximate 95% confidence intervals and hypothesis tests based on a *t* distribution; otherwise, the between-within DOF approximation for *t*-based inference by Schluchter and Elashoff (1990) was used.
The small-sample correction to the sandwich variance estimator was based on the approximation by Mancl and DeRouen (2001).
The regression coefficients of primary interest are β_1 and β_{11}; specifically, the null hypothesis $H_0: \beta_1 + \beta_{11} = 0$ corresponds to no treatment difference in the mean change from baseline at week 48 adjusting for the other covariates; this is similar to testing H_0: Treament Difference in LS Means at Week $48 = 0$ (i.e., the marginal mean change from baseline is the same in each arm).

Hypothesis of Interest: H_0: Difference in LS Means at Week $48 = 0$ (i.e., the marginal mean change from baseline at week 48 is the same in each arm)

Operating Characteristics
Results based on 100,000 Simulated Data Sets

Model Number: Model Description	Coefficient	Average	Bias	Average SE	95% Coverage	Difference in LS Means (SE) Week 48	Power of Test H_0: Difference in LS Means at Week $48 = 0$ vs H_a: Difference in LS Means at Week $48 \neq 0$
M0: MMRM Model, UN Covariance Correctly Specified Kenward-Roger Adjusted SEs *T*-based CIs and Tests	β_1	0.8334	-0.0003	0.8721	0.9502	Treatment: -2.3453 Placebo: -3.1355 Placebo – Treatment: -0.7902 (1.2238)	0.0973
	β_{11}	-0.0432	-0.0055	1.1117	0.9495		
M1: MMRM Model, CS Model-based SEs *T*-based CIs and Tests	β_1	0.8331	-0.0006	1.1548	0.9902	Treatment: -2.3456 Placebo: -3.1354 Placebo – Treatment: -0.7898 (1.1548)	0.1227
	β_{11}	-0.0433	-0.0056	1.0196	0.9250		

Appendix Table 2A. (continued)

M2: MMRM Model, CS **Sandwich Estimator (Robust SEs)** ***T*-based CIs and Tests**	β_1	0.8331	-0.0006	0.8630	0.9463	Treatment: -2.3456 Placebo: -3.1354 Placebo – Treatment: -0.7898 (1.1977)	0.1100
	β_{11}	-0.0433	-0.0056	1.1000	0.9437		
M3: MMRM Model, CS **Sandwich Estimator (Robust SEs)** **Small-sample Correction** ***T*-based CIs and Tests**	β_1	0.8331	-0.0006	0.9083	0.9572	Treatment: -2.3456 Placebo: -3.1354 Placebo – Treatment: -0.7898 (1.2560)	0.0941
	β_{11}	-0.0433	-0.0056	1.1262	0.9492		
M4: GEE Model, CS **Model-based SEs** ***Z*-based CIs and Tests**	β_1	0.8331	-0.0006	1.1445	0.9884	Treatment: -2.3456 Placebo: -3.1354 Placebo – Treatment: -0.7898 (1.1446)	0.1273
	β_{11}	-0.0433	-0.0056	1.0357	0.9287		
M5: GEE Model, CS **Sandwich Estimator (Robust SEs)** ***Z*-based CIs and Tests**	β_1	0.8331	-0.0006	0.8629	0.9424	Treatment: -2.3456 Placebo: -3.1354 Placebo – Treatment: -0.7898 (1.1977)	0.1113
	β_{11}	-0.0433	-0.0056	1.1000	0.9429		
M6: GEE Model, CS **Sandwich Estimator (Robust SEs)** **Small-sample Correction** ***Z*-based CIs and Tests**	β_1	0.8331	-0.0006	0.9082	0.9544	Treatment: -2.3456 Placebo: -3.1354 Placebo – Treatment: -0.7898 (1.2560)	0.0950
	β_{11}	-0.0433	-0.0056	1.1261	0.9484		
M7: GEE Model, CS **Model-based SEs** ***T*-based CIs and Tests**	β_1	0.8331	-0.0006	1.1445	0.9897	Treatment: -2.3456 Placebo: -3.1354 Placebo – Treatment: -0.7898 (1.1446)	0.1259
	β_{11}	-0.0433	-0.0056	1.0357	0.9296		
M8: GEE Model, CS **Sandwich Estimator (Robust SEs)** ***T*-based CIs and Tests**	β_1	0.8331	-0.0006	0.8629	0.9462	Treatment: -2.3456 Placebo: -3.1354 Placebo – Treatment: -0.7898 (1.1977)	0.1100
	β_{11}	-0.0433	-0.0056	1.1000	0.9436		
M9: GEE Model, CS **Sandwich Estimator (Robust SEs)** **Small-sample Correction** ***T*-based CIs and Tests**	β_1	0.8331	-0.0006	0.9082	0.9572	Treatment: -2.3428 Placebo: -3.1354 Placebo – Treatment: -0.7898 (1.2560)	0.0941
	β_{11}	-0.0433	-0.0056	1.1261	0.9492		

Appendix Table 2B. Simulation study and comprehensive results for study UX023-CL301 (Pediatrics).

Study UX023-CL301
N = 61, Two Arm Study

Data Generating Model

MMRM Model: $E(Y_i|X_i) = \beta_0 + \beta_1 I(\mathrm{G}=2) + \beta_2 I(\mathrm{W}=64) + \beta_3 I(\mathrm{G}=2)\times I(\mathrm{W}=64) + \gamma_1 \mathrm{TOTSCORE} + \gamma_2 I(\mathrm{Age}=2)$

Repeated measurements taken on $i = 1, \ldots, 61$ participants (32 conventional therapy; 29 burosumab) at the same set of 2 post-baseline occasions.
Y_i = Change from baseline.
I represents an indicator variable.
G = Group; W = Week; TOTSCORE = Baseline Total Rickets Severity Score; Reference categories for variables G, W, and Age were burosumab, week 40, and < 5 years, respectively.
Unstructured (UN) covariance structure was assumed.
The estimates of the regression coefficients and variance / covariance matrix from fitting the MMRM model to the trial data were used as the true values.
The covariates TOTSCORE and Age were generated from a distribution reasonably consistent with the trial data: a truncated *N*(3.18, 1.057) and 57% ≥ 5 years old.
100,000 data sets were generated; on average, in any given simulated data set, 0.5% of the data were missing completely at random.
When the Kenward-Roger adjusted standard errors were used, Satterthwaite-type degrees of freedom (DOF) were computed based on this adjustment and used to construct approximate 95% confidence intervals and hypothesis tests based on a *t* distribution; otherwise, the between-within DOF approximation for *t*-based inference by Schluchter and Elashoff (1990) was used.
The small-sample correction to the sandwich variance estimator was based on the approximation by Mancl and DeRouen (2001).
The regression coefficient of primary interest, β_1, corresponds to the treatment difference in the mean change from baseline at week 40 adjusting for the other covariates; this is similar to testing H_0: Treament Difference in LS Means at Week $40 = 0$ (i.e., the marginal mean change from baseline is the same in each arm).

Primary Hypothesis of Interest: H_0: Difference in LS Means at Week $40 = 0$ (i.e., the marginal mean change from baseline at week 40 is the same in each arm)

Operating Characteristics
Results based on 100,000 Simulated Data Sets

	Coefficient	Average	Bias	Average SE	95% Coverage	Difference in LS Means (SE) Week 40	Power of Test H_0: Difference in LS Means at Week $40 = 0$ vs H_a: Difference in LS Means at Week $40 \neq 0$
M0: MMRM Model, UN Covariance Correctly Specified Kenward-Roger Adjusted SEs *T*-based CIs and Tests	β_1	1.3578	-0.0008	0.2011	0.9501	Burosumab: -2.0793 Conventional Therapy: -0.7215 Burosumab – Conv Ther: -1.3578 (0.2011)	1.00
M1: MMRM Model, CS Model-based SEs *T*-based CIs and Tests	β_1	1.3578	-0.0008	0.1996	0.9494	Burosumab: -2.0793 Conventional Therapy: -0.7215 Burosumab – Conv Ther: -1.3578 (0.1996)	1.00
M2: MMRM Model, CS Sandwich Estimator (Robust SEs) *T*-based CIs and Tests	β_1	1.3578	-0.0008	0.1945	0.9425	Burosumab: -2.0793 Conventional Therapy: -0.7215 Burosumab – Conv Ther: -1.3578 (0.1945)	1.00

Appendix Table 2B. (continued)

M3: MMRM Model, CS **Sandwich Estimator (Robust SEs)** **Small-sample Correction** ***T*-based CIs and Tests**	β_1	1.3578	-0.0008	0.2076	0.9563	Burosumab: -2.0793 Conventional Therapy: -0.7215 Burosumab – Conv Ther: -1.3578 (0.2076)	1.00
M4: GEE Model, CS **Model-based SEs** ***Z*-based CIs and Tests**	β_1	1.3578	-0.0008	0.1984	0.9430	Burosumab: -2.0793 Conventional Therapy: -0.7215 Burosumab – Conv Ther: -1.3578 (0.1984)	1.00
M5: GEE Model, CS **Sandwich Estimator (Robust SEs)** ***Z*-based CIs and Tests**	β_1	1.3578	-0.0008	0.1945	0.9375	Burosumab: -2.0793 Conventional Therapy: -0.7215 Burosumab – Conv Ther: -1.3578 (0.1945)	1.00
M6: GEE Model, CS **Sandwich Estimator (Robust SEs)** **Small-sample Correction** ***Z*-based CIs and Tests**	β_1	1.3578	-0.0008	0.2076	0.9520	Burosumab: -2.0793 Conventional Therapy: -0.7215 Burosumab – Conv Ther: -1.3578 (0.2076)	1.00
M7: GEE Model, CS **Model-based SEs** ***T*-based CIs and Tests**	β_1	1.3578	-0.0008	0.1984	0.9480	Burosumab: -2.0793 Conventional Therapy: -0.7215 Burosumab – Conv Ther: -1.3578 (0.1984)	1.00
M8: GEE Model, CS **Sandwich Estimator (Robust SEs)** ***T*-based CIs and Tests**	β_1	1.3578	-0.0008	0.1945	0.9424	Burosumab: -2.0793 Conventional Therapy: -0.7215 Burosumab – Conv Ther: -1.3578 (0.1945)	1.00
M9: GEE Model, CS **Sandwich Estimator (Robust SEs)** **Small-sample Correction** ***T*-based CIs and Tests**	β_1	1.3578	-0.0008	0.2076	0.9563	Burosumab: -2.0793 Conventional Therapy: -0.7215 Burosumab – Conv Ther: -1.3578 (0.2076)	1.00

Appendix Table 2C. Simulation study and comprehensive results for study UX023-CL304 (Adults).

Study UX023-CL304
N = 14, Single Arm Study

Data Generating Model

MMRM Model: $E(Y_i|X_i) = \beta_0 + \beta_1 I(\mathrm{W} = 12) + \beta_2 I(\mathrm{W} = 24) + \beta_3 I(\mathrm{W} = 36) + \beta_4 I(\mathrm{W} = 48) + \gamma_1 \mathrm{Base}$

Repeated measurements taken on $i = 1, \ldots, 14$ participants at the same set of 5 post-baseline occasions.
Y_i = Change from baseline in alkaline phosphatase.
I represents an indicator variable.
W = Week; Base = Baseline value; Reference category for the variable W was study week 4.
Unstructured (UN) covariance structure was assumed.
The estimates of the regression coefficients and variance / covariance matrix from fitting the MMRM model to the trial data were used as the true values.
The covariate Base was generated from a distribution reasonably consistent with the trial data: a truncated *N*(113.14, 51.600).
500,000 data sets were generated; on average, in any given simulated data set, 1.2% of the data were missing completely at random.
When the Kenward-Roger adjusted standard errors were used, Satterthwaite-type degrees of freedom were computed based on this adjustment and used to construct approximate 95% confidence intervals and hypothesis tests based on a *t* distribution; otherwise, the between-within DOF approximation for *t*-based inference by Schluchter and Elashoff (1990) was used.
The small-sample correction to the sandwich variance estimator was based on the approximation by Mancl and DeRouen (2001).
The regression coefficients of primary interest are β_0 and β_4; specifically, the null hypothesis H_0: $\beta_0 + \beta_4 = 0$ corresponds to no difference in the mean change from baseline at week 48 adjusting for the baseline value. This is similar to testing H_0: LS Mean at Week $48 = 0$ (i.e., the marginal mean change from baseline equals zero).

Primary Hypothesis of Interest: H_0: LS Means at Week $48 = 0$ (i.e., the estimated marginal mean change from baseline at week 48 equals zero)

Operating Characteristics
Results based on 500,000 Simulated Data Sets

	Coefficient	Average	Bias	Average SE	95% Coverage	LS Means (SE) Week 48	Power of Test H_0: LS Means at Week $48 = 0$ vs H_a: LS Means at Week $48 \neq 0$
M0: MMRM Model, UN Covariance Correctly Specified Kenward-Roger Adjusted SEs *T*-based CIs and Tests	β_0	24.2090	-0.0040	10.7878	0.9200	1.1379 (6.5140)	0.0529
	β_1	-3.0681	0.0033	5.0137	0.9503		
	β_2	-17.2791	0.0066	6.9124	0.9496		
	β_3	-20.3500	0.0071	8.4653	0.9497		
	β_4	-20.1646	0.0085	9.5228	0.9500		

Appendix Table 2C. (continued)

M1: MMRM Model, CS **Model-based SEs** ***T*-based CIs and Tests**	β_0	24.1961	-0.0169	12.1235	0.9477	1.1363 (5.8538)	0.0902
	β_1	-3.0694	0.0020	6.4422	0.9810		
	β_2	-17.2801	0.0056	6.4424	0.9219		
	β_3	-20.3496	0.0075	6.4425	0.8544		
	β_4	-20.1670	0.0061	6.4423	0.8071		
M2: MMRM Model, CS **Sandwich Estimator (Robust SEs)** ***T*-based CIs and Tests**	β_0	24.1961	-0.0169	11.0114	0.9124	1.1363 (6.1476)	0.0861
	β_1	-3.0694	0.0020	4.8793	0.9258		
	β_2	-17.2801	0.0056	6.6747	0.9243		
	β_3	-20.3496	0.0075	8.1734	0.9243		
	β_4	-20.1670	0.0061	9.1938	0.9244		
M3: MMRM Model, CS **Sandwich Estimator (Robust SEs)** **Small-sample Correction** ***T*-based CIs and Tests**	β_0	24.1961	-0.0169	13.9071	0.9554	1.1363 (6.9064)	0.0575
	β_1	-3.0694	0.0020	5.2634	0.9436		
	β_2	-17.2801	0.0056	7.1991	0.9419		
	β_3	-20.3496	0.0075	8.8154	0.9421		
	β_4	-20.1670	0.0061	9.9163	0.9423		
M4: GEE Model, CS **Model-based SEs** ***Z*-based CIs and Tests**	β_0	24.1961	-0.0169	11.6849	0.9130	1.1363 (5.7849)	0.1015
	β_1	-3.0693	0.0021	6.5826	0.9812		
	β_2	-17.2800	0.0057	6.5829	0.9219		
	β_3	-20.3496	0.0075	6.5829	0.8543		
	β_4	-20.1669	0.0062	6.5827	0.8067		
M5: GEE Model, CS **Sandwich Estimator (Robust SEs)** ***Z*-based CIs and Tests**	β_0	24.1961	-0.0169	11.0090	0.8814	1.1363 (6.1473)	0.0929
	β_1	-3.0693	0.0021	4.8795	0.9195		
	β_2	-17.2800	0.0057	6.6745	0.9179		
	β_3	-20.3496	0.0075	8.1729	0.9180		
	β_4	-20.1669	0.0062	9.1940	0.9183		

Appendix Table 2C. (continued)

M6: GEE Model, CS **Sandwich Estimator (Robust SEs)** **Small-sample Correction** ***Z*-based CIs and Tests**	β_0	24.1961	-0.0170	13.9045	0.9356	1.1363 (6.9062)	0.0627
	β_1	-3.0693	0.0021	5.2639	0.9383		
	β_2	-17.2800	0.0057	7.1990	0.9366		
	β_3	-20.3496	0.0075	8.8150	0.9367		
	β_4	-20.1669	0.0062	9.9165	0.9371		
M7: GEE Model, CS **Model-based SEs** ***T*-based CIs and Tests**	β_0	24.1961	-0.0169	11.6849	0.9400	1.1363 (5.7849)	0.0938
	β_1	-3.0693	0.0021	6.5826	0.9837		
	β_2	-17.2800	0.0057	6.5829	0.9284		
	β_3	-20.3496	0.0075	6.5829	0.8634		
	β_4	-20.1669	0.0062	6.5827	0.8168		
M8: GEE Model, CS **Sandwich Estimator (Robust SEs)** ***T*-based CIs and Tests**	β_0	24.1961	-0.0169	11.0090	0.9124	1.1363 (6.1473)	0.0862
	β_1	-3.0693	0.0021	4.8795	0.9258		
	β_2	-17.2800	0.0057	6.6745	0.9243		
	β_3	-20.3496	0.0075	8.1729	0.9243		
	β_4	-20.1669	0.0062	9.1940	0.9244		
M9: GEE Model, CS **Sandwich Estimator (Robust SEs)** **Small-sample Correction** ***T*-based CIs and Tests**	β_0	24.1961	-0.0169	13.9045	0.9553	1.1363 (6.9062)	0.0575
	β_1	-3.0693	0.0021	5.2639	0.9436		
	β_2	-17.2800	0.0057	7.1990	0.9418		
	β_3	-20.3496	0.0075	8.8150	0.9420		
	β_4	-20.1669	0.0062	9.9165	0.9424		

David Zahrieh. Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN.
E-mail address: zahrieh.david@mayo.edu

Yi Wang. Department of Data Sciences & Development Strategy, Ultragenyx Pharmaceutical, Novato, CA

Jennifer Le-Rademacher. Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN

Tony Koutsoukos. Department of Data Sciences & Development Strategy, Ultragenyx Pharmaceutical, Novato, CA