Examining the Feasibility and Functionality of Common Analysis Methods for Longitudinal Data with Binary Outcomes

Isabella Zaniletti

Mayo Clinic

Amylou C. Dueck

Mayo Clinic

Todd A. DeWees (✉ DeWees.Todd@mayo.edu)

Mayo Clinic

Research Article

Keywords: Power analysis, binomial distribution, longitudinal data, dichotomous outcome

Posted Date: February 15th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1299793/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background

Previous literature often indicates that binary longitudinal data, such as responder analysis with patient-reported outcomes to evaluate treatment risks and benefits in cancer randomized controlled trials, need complex statistical approaches for a thorough evaluation of the treatment effect and consequently elaborated clinical design and power calculation for prospective trials. This simulated analysis aims to evaluate six common statistical techniques used for binary longitudinal data to explore the feasibility and overall functionality of these methods in comparison to more complex modeling techniques.

Methods

We simulated data to represent a typical clinical trial design, with dichotomous outcome, and treatment vs control group at three waves T1, T2, and T3. We evaluated five sample size conditions ranging from small to large with 1,000 replicates each. We then evaluated results from six statistical techniques (Generalized linear model [GLM] with generalized estimating equations [GEE], Generalized linear mixed model [GLMM], Logistic regression, Cochran-Mantel-Haenszel [CMH], Chi-square, and Fisher's Exact test) and their ability in detecting difference between treatment groups under different conditions.

Results

While less elaborated models always achieved convergence, the convergence rate of GLMM with unstructured covariance matrix ranged from 7–43%. The proportion of detected significant differences between control and treatment expectedly increased with the increasing of the sample size but remained somewhat similar across modeled outcome and sample size scenarios.

Conclusions

Overall results indicated that, with careful consideration, straightforward modelling approaches which require less assumptions in design and power calculation are sufficient to ensure meaningful, adequately powered, evaluation of treatment effect.

Introduction

The increase in the use of longitudinally collected patient-reported outcomes (PROs) to evaluate treatment risks and benefits in cancer randomized controlled trials (RCTs) has put emphasis on the need to evaluate appropriate approaches to support a trial design and thorough statistical analyses. In the context of clinical trials where the exposure is time-invariant, the aim of study involving repeated measurements is typically the change in response of treatment and control. Although several approaches to data analysis of longitudinal RCT's have been proposed, there is not a gold standard. The Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life (SISAQOL) Endpoints Data Consortium provides recommendation on several aspects of PROs analysis, from how to develop a taxonomy of research objectives, to how to handle missing values (1). The proposed statistical methods for answering the broad selection of research questions within RCTs vary from *t*-tests, Cox regression, Analysis of Variance (ANOVA), Analysis of Covariance (ANCOVA), and mixed models with further considerations of adjustments, clusters, and interactions (1). Although many approaches to handle data analysis for longitudinal data in clinical trials exist, power and sample size methods are available only for a limited class of these models. Moreover, even if the use of longitudinal data often provides increased statistical power towards examining causal effects and treatment differences, the complexity of these designs along with the underlying correlation structure can impact the application of the planned statistical analysis. In addition to the challenging task of identifying the method that may 'best' answer specific research questions, we face with the apparent need of ensuring, within complex data structures, the appropriate sample size through statistical power estimation to meaningfully evaluate the analysis results.

For the last three decades, several method publications have been emphasizing the need of sophisticated techniques to power studies involving longitudinal data, underlying the possible effect of within-subject correlations, repeated measurements, and missing data on the effect-size estimation. Rochon (2) adapted the Liang and Zeger's approach of performing sample size calculation for repeated measure experiment, using the generalized estimating equations (GEE) model with a model-based covariance, while Muller focused their analysis on general linear multivariate models (3). In their extensive work on sample size estimation for longitudinal designs with attrition, Hedeker, Gibbons and Watnaux (4) derived several formulas under assumptions of compound symmetry, first-order autoregressive, and non-stationary random-effects structure, applicable to a wide variety of models. Tu developed an alternative approach to GEE and mixed-effect models, which derives the power function on the asymptotic distribution of the model estimate. This method attempts to improve the limitations for clustered approaches, however, it requires a strong assumption of a common constant cluster size across clusters (5). Many authors recommended sensitivity analysis to assess sample size requirements through variation of the non-centrality parameters (3, 4) and others recommended inflation of the resulting sample size

to protect from deviations in the assumptions (2). More recent work focuses on implementing the formulation for sample size calculation to compare difference among groups over time using GEE method that account for missing patterns, correlation structures and unbalance designs (6),(7), and extend the approach to multivariate analysis (8). Although these techniques aid in power calculations for the planned analyses, the validity of their results are based on strong assumptions of the parameter estimates and covariance structure. Often, the failure to meet these assumptions results in non-estimable model parameters forcing the use of alternate modeling techniques. Furthermore, all proposed approaches seem to require a rather intense effort in terms of evaluation and coding, while model convergence may not be achieved anyway. Since the consensus around the best approach remains open (1) a question arises, could we simplistically reduce the power calculation process to a less computationally intensive approach, without compromising statistical integrity? In this paper we provide a thorough analysis of six common approaches to longitudinal data for binomial outcomes to verify whether a straightforward and more direct approach to power analysis is plausible. We use simulated data to illustrate differences, similarities, and feasibility of the following six techniques under selected parameter combinations: 1) Generalized linear model (GLM) with generalized estimating equations (GEE); 2) Generalized linear mixed model (GLMM); 3) Logistic regression; 4) Cochran-Mantel-Haenszel (CMH) 5) Chi-square, and 6) Fisher's Exact test.

Methods

Data was simulated to represent a typical clinical trial scenario with two arms (A0=control and A1=treatment) and three wave measurements, T1, T2, and T3. Multivariate normal data was generated with specification of population mean vector and covariance matrix. The arm assignment was generated using the Bernoulli distribution, and the means at time 2 and 3 for control and treatment arms were finalized according to Table 1. The outcome of interest was then dichotomized based on location of the generated value being less than or greater than or equal to the generated population mean. Samples of sizes 50, 300, 500 and 1,000 where each replicated 1,000 times.

Table 1
Summary of simulation parameters

Waves	T1	T2	T3
arm = RAND('BINOMIAL',0.50,1)			
Control mean	60	55	50
Treatment mean	60	50	40
SD	20	20	20
Correlation	T1-T2 = 0.4	T2-T3 = 0.7	T3-T1 = 0.4
Outcome (Event = 1)	if T1 $\geq$ 60	if T2 $\geq$ 60	if T3 $\geq$ 60

For each replicated sample, of any size, the hypothetical event was first modeled with a 1) GLM for correlated binary data and a GEE approach (9). The correlation structure was set to unstructured, to allow distinct correlation elements to differ, binomial distribution with logit link was used, along with a repeated statement on subjects. The outcome was modeled in terms of arm, wave, and their interaction. Contrast statements were utilized for parameter estimate evaluation. We then analyzed with a 2) GLMM with wave as random effect, under two different covariance matrices: unstructured and first-order autoregressive, AR(1). While the unstructured covariance structure estimates unique correlations for each time point pairs, the first-order autoregressive, matrix is calculated based on the previous wave and it considers correlations to be highest for adjacent times, systematically decreasing with increasing distance between time points. For both GEE and GLMM, the outcome was first modeled in terms of arm and time (T1, T2, and T3), and subsequently only accounting for T2 and T3 and adjusting for baseline. Moreover, baseline was considered as both a dichotomous and then continuous value. Estimates and 3) Wald's confidence intervals were obtained for the logistic regression in which the dichotomous outcome at T3 is regressed on arm, using baseline as covariate, both as dichotomous and continuous variable. 5) Chi-square and 6) Fisher's exact were used to produce the test for association between arm and outcome, both at third wave, while 4) CMH was used to evaluate the association between outcomes at all non-baseline waves and arm controlling for baseline. We first graphically displayed and evaluated convergence percentages of each model. Then we analyzed the ability of each model to detect significance in the treatment as a measure of power. The count of times each model detected significance was modeled using a Poisson regression. The logarithm of the number of times each model converged was considered as offset for the model. We used the most complex of the models, the GLMM for the outcome at T3 with baseline (as continuous) adjustment and unstructured covariance, as the referent group, and we provided odds ratios for the comparison of the convergence of each model with the referent. All analyses were conducted in SAS v9.4 (Cary, NC), and results evaluated at the significance level set at 0.05.

Results

We simulated data to compare six statistical approaches (Table 2) to the evaluation of a treatment effect within a longitudinal analysis setting with a binary outcome. We first looked at model convergence. All models, except for the GLMM models with unstructured covariance matrix,

achieved convergence 100% of the time for all replicates, sample sizes, and conditions (Figure 1). For the largest sample (n=1,000), GLMM with unstructured covariance matrix converged only 7% of the time when modeling the change from baseline to T3 or the outcome at T3 directly. Convergence for the GLMM increased as the sample size decreased. The maximum GLMM convergence of 43% was achieved for the smallest sample size (n=50). The more consistent convergence of GLMM with unstructured covariance occurred when modeling the outcome at T3, adjusting for baseline as a continuous measure.

Table 2. Model specifications

Approach	Model	Time of Assessment	Predictor	Adjustor	Covariance Structure
CMH	Cochran-Mantel-Haenszel	T2,T3	Arm	baseline	n/a
Chi-Square	Chi-square	T3	Arm	n/a	n/a
GENMOD→Chg Base to T3 → A1-A0	Generalized Linear with GEE	T1,T2,T3	Arm wave	none	Unstructured
GENMOD→T=3 → A1-A0	Generalized Linear with GEE	T2,T3	Arm wave	none	Unstructured
GENMOD→T=3 Cntrl 0/1 → A1-A0	Generalized Linear with GEE	T2,T3	Arm wave	baseline (continuous)	Unstructured
GENMOD→T=3 Cntrl Cont → A1-A0	Generalized Linear with GEE	T2,T3	Arm wave	baseline (dichotomous)	Unstructured
GLIMMIX UN→Chg Base to T3 → A1-A0	Generalized Linear Mixed	T1,T2,T3	Arm wave	none	Unstructured
GLIMMIX UN→T=3 → A1-A0	Generalized Linear Mixed	T2,T3	Arm wave	none	Unstructured
GLIMMIX UN→T=3 Cntrl 0/1 → A1-A0	Generalized Linear Mixed	T2,T3	Arm wave	baseline (continuous)	Unstructured
GLIMMIX UN→T=3 Cntrl Cont → A1-A0	Generalized Linear Mixed	T2,T3	Arm wave	baseline (dichotomous)	Unstructured
GLIMMIX UN→Chg Base to T3 → A1-A0	Generalized Linear Mixed	T1,T2,T3	Arm wave	none	AR(1)
GLIMMIX UN→T=3 → A1-A0	Generalized Linear Mixed	T2,T3	Arm wave	none	AR(1)
GLIMMIX UN→T=3 Cntrl 0/1 → A1-A0	Generalized Linear Mixed	T2,T3	Arm wave	baseline (continuous)	AR(1)
GLIMMIX UN→T=3 Cntrl Cont → A1-A0	Generalized Linear Mixed	T2,T3	Arm wave	baseline (dichotomous)	AR(1)
Logistic→0/1 Baseline	Logistic Regression	T3	Arm	baseline (continuous)	n/a
Logistic→Cont. Base	Logistic Regression			baseline (dichotomous)	n/a
XP2_FISH,Two-sided Pr ≤ P	Fisher's Exact	T3	Arm	n/a	n/a

In Table 3 we assessed the proportion of detected significant difference between control and treatment of each model when compared to the referent model. A significant OR>1 indicates that the model is more likely to detect a statistically significant difference between groups than the referent. Conversely, a significant OR<1 indicates that the model is less likely to detect the difference than the GLMM. When we assessed the odds ratios across models and sample sizes we found an expected larger proportion of detected group difference across models for larger sample size. This measure of power remained similar across methods (Table 3). In fact, when modeling the rate of successfully detecting treatment effect, by sample size and across the different approaches, we found only few inconsistent results across models and sample size scenarios. Only for sample size n=50, CMH showed different results than the referent model, detecting approximately 58% more significance in the treatment effect than the GLMM for the outcome at T3 with continuous baseline adjustment and unstructured covariance. For the smallest sample size, the GEE approach and the GLMM for the change were more conservative than the referent model, estimating 35% and 46% less cases respectively. The GLMM models for the outcome at T3 with first-order autoregressive covariance matrix and baseline adjustment (as both dichotomous and continuous) also detected 30% less treatment effect than the GLMM with unstructured covariance matrix and baseline (as continuous)

adjustment. Finally, the GEE approach and the GLMM for the change showed significant difference in detecting difference compare to the referent also for sample sizes n=150 (OR=0.76 with 95%CI 0.66,0.89 p<.01 and OR=0.62 with 95% CI 0.53,0.72 p<.01 respectively) and for sample size n=300 (OR=0.85 with 95% CI 0.74,0.96 p=.01 and OR=0.77 with 95% CI 0.68,0.88 p<.01 respectively (Table 3).

Table 3
Summary of Poisson regression for the detection rate of statistically significant treatment effect

Approach*	n=50		n=150		n=300		n=500		n=1000	
	OR + 95% CI	p	OR + 95% CI	p	OR + 95% CI	p	OR + 95% CI	p	OR + 95% CI	p
CMH	1.58[1.23,2.02]	<.01	1.01[0.87,1.16]	0.94	0.97[0.86,1.1]	0.64	0.98[0.87,1.11]	0.73	1[0.88,1.14]	0.99
Chi-Square	1.25[0.97,1.62]	0.08	1[0.87,1.16]	0.99	0.97[0.85,1.1]	0.63	0.99[0.88,1.12]	0.89	1[0.88,1.14]	0.99
GENMOD->Chg Base to T3 -> A1-A0	0.75[0.57,0.98]	0.04	0.76[0.66,0.89]	<.01	0.85[0.74,0.96]	0.01	0.95[0.84,1.07]	0.37	1[0.88,1.14]	0.98
GENMOD->T=3 -> A1-A0	1.01[0.78,1.31]	0.94	0.99[0.85,1.14]	0.86	0.97[0.85,1.1]	0.60	0.99[0.88,1.12]	0.89	1[0.88,1.14]	0.99
GENMOD->T=3 Cntrl 0/1 -> A1-A0	1[0.77,1.3]	0.99	0.99[0.86,1.15]	0.92	0.98[0.86,1.11]	0.74	0.99[0.88,1.12]	0.92	1[0.88,1.14]	0.99
GENMOD->T=3 Cntrl Cont -> A1-A0	1.11[0.86,1.44]	0.42	1.01[0.87,1.17]	0.89	1[0.88,1.13]	0.94	1[0.88,1.13]	0.95	1[0.88,1.14]	0.99
GLIMMIX UN->Chg Base to T3 -> A1-A0	0.68[0.46,1.02]	0.06	0.81[0.66,1]	0.05	0.91[0.76,1.1]	0.33	0.95[0.79,1.15]	0.61	1[0.77,1.3]	0.99
GLIMMIX UN->T=3 -> A1-A0	1[0.7,1.42]	0.99	0.96[0.79,1.17]	0.71	0.98[0.82,1.17]	0.78	0.99[0.83,1.2]	0.95	1[0.77,1.3]	0.99
GLIMMIX UN->T=3 Cntrl 0/1 -> A1-A0	0.95[0.7,1.31]	0.76	0.98[0.82,1.17]	0.82	0.99[0.85,1.16]	0.92	0.99[0.85,1.17]	0.95	1[0.82,1.22]	0.99
GLIMMIX->Chg Base to T3 -> A1-A0	0.54[0.4,0.72]	<.01	0.62[0.53,0.72]	<.01	0.77[0.68,0.88]	<.01	0.91[0.8,1.03]	0.13	1[0.88,1.14]	0.98
GLIMMIX->T=3 -> A1-A0	0.77[0.58,1.01]	0.06	0.89[0.77,1.03]	0.11	0.94[0.83,1.06]	0.32	0.99[0.87,1.12]	0.84	1[0.88,1.14]	0.99
GLIMMIX->T=3 Cntrl 0/1 -> A1-A0	0.7[0.53,0.93]	0.01	0.89[0.77,1.03]	0.11	0.94[0.83,1.07]	0.35	0.98[0.87,1.11]	0.81	1[0.88,1.14]	0.99
GLIMMIX->T=3 Cntrl Cont -> A1-A0	0.7[0.53,0.93]	0.01	0.91[0.78,1.05]	0.20	0.96[0.84,1.08]	0.48	0.98[0.87,1.11]	0.81	1[0.88,1.14]	0.99
Logistic->0/1 Baseline	0.97[0.74,1.26]	0.80	1[0.86,1.15]	0.97	0.99[0.87,1.12]	0.84	0.99[0.88,1.12]	0.89	1[0.88,1.14]	0.99
Logistic->Cont. Base	1.03[0.79,1.33]	0.85	1.01[0.87,1.16]	0.91	0.99[0.87,1.12]	0.88	0.99[0.88,1.12]	0.91	1[0.88,1.14]	0.99
XP2_FISH,Two-sided Pr <= P	0.93[0.71,1.21]	0.59	0.92[0.79,1.06]	0.25	0.95[0.84,1.08]	0.44	0.99[0.88,1.12]	0.88	1[0.88,1.14]	0.99
*Referent model="GLIMMIX UN->T=3 Cntrl Cont -> A1-A0"										

Conclusion

The primary hypothesis was to identify if and when more complex methods can improve statistical analyses over less intensive models to analyze the effect of a predictor in the context of binary longitudinal analysis, and if such information could be used to simplify the initial task of designing and powering clinical studies. We sought for answers fitting several models on simulated data and under different parameter combinations. Our findings indicated that, with few exceptions, results from the approaches considered do not significantly differ from each other. Besides CMH, which demonstrated to be slightly liberal for small samples, only the GLMM with first-order autoregressive covariance structure with baseline adjustment differ from the referent when model outcome at T3. Modeling the change directly appears to be more conservative than

modeling the outcome at T3, and this is consistent across sample sizes. Chi-square test, Fisher's Exact, logistic regression, CMH and most of the GEE and GLMM demonstrate similar results, with small variation across methods in terms of power as shown by the number of statistically significant results. According to the analysis results any of the less computationally intense models were a dependable choice. When considering the trade-off between model simplicity and flexibility, baseline adjusted logistic regression seem to be the desirable model as, compared to CMH, Chi-squares and Fisher's Exact, it allows for covariate adjustment and models selection, it provides odds ratios and estimated confidence intervals, easily achieves 100% convergence, all remaining of fairly simple application and interpretation. More complex models, such as GLM and GLMM, highly depend on strong assumptions that in many cases are neither met nor verifiable. As showed in our analysis, small changes in the assumptions for these types of models influence the overall performance, and even if we are confident in such assumptions, in many cases these models fail to converge, forcing the researcher to look for alternative and less demanding approaches. In our example, the convergence of the GLMM counterintuitively decreases as the sample size increases. Since typically fitting errors are not only based on sample size and sparseness of the data, but most of all by the increased complexity of the model itself, it is reasonable to assert that GLMM is not the best choice (at least in this form, and for our data).

We recognize several limitations in our study. The major issue concerns the generalizability of our findings. The evaluation of the presented analysis results remains meaningful within the simulation parameters of our data. Moreover, additional analysis on more complex simulated data and starting parameters is needed to sustain more robust conclusion. Although vary sample sizes with several replicates were evaluated, additional simulations with multiple choice of distribution parameters for both outcome and arm might be beneficial. We also understand that more attentive evaluation of the parameters within the GLMM approach could have led to an empirically sound model that might have achieved, maybe, more convergence. This said, we believe we made a good point of the laborious effort these models require in order to do so.

After an investigation into the trade-off between effort required, convergence achieved, and expected results, methods with fewer assumptions and model complexity provide similar results to their counterparts without reduction in statistical power or statistical rigor.

Declarations

Ethics approval and consent to participate: Not Applicable

Consent for publication: Not Applicable

Availability of data and materials: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests: The authors declare that they have no competing interests.

Funding: No funding received for the manuscript

Author's contributions: IZ conceptualized and design the work, simulated and analyzed the data, interpreted the data, drafted and revised the manuscript. ACD conceptualized the work, contribute to data interpretation and manuscript. TAD conceptualized the work, contribute to data interpretation and manuscript. All authors read and approved the final manuscript.

Acknowledgements: Not Applicable

References

1. Coens C, Pe M, Dueck AC, Sloan J, Basch E, Calvert M, et al. International standards for the analysis of quality-of-life and patient-reported outcome endpoints in cancer randomised controlled trials: recommendations of the SISAQOL Consortium. *Lancet Oncol.* 2020;21(2):e83-e96.
2. Rochon J. Application of GEE procedures for sample size calculations in repeated measures experiments. *Statistics in medicine.* 1998;17(14):1643-58.
3. Muller KE, Lavange LM, Ramey SL, Ramey CT. Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association.* 1992;87(420):1209-26.
4. Hedeker D, Gibbons RD, Waternaux C. Sample size estimation for longitudinal designs with attrition: comparing time-related contrasts between two groups. *Journal of Educational and Behavioral Statistics.* 1999;24(1):70-93.
5. Tu XM, Kowalski J, Zhang J, Lynch K, Crits-Christoph P. Power analyses for longitudinal trials and other clustered designs. *Statistics in medicine.* 2004;23(18):2799-815.
6. Lou Y, Cao J, Zhang S, Ahn C. Sample size estimation for a two-group comparison of repeated count outcomes using GEE. *Communications in Statistics-Theory and Methods.* 2017;46(14):6743-53.
7. Wang J, Zhang S, Ahn C. Sample size calculation for comparing time-averaged responses in <italic>K</italic>-group repeated binary outcomes. *CSAM.* 2018;25(3):321-8.

8. Liu G, Liang K-Y. Sample size calculations for studies with correlated observations. Biometrics. 1997;937-47.

9. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73(1):13-22.

Figures

Figure 1

Distribution of model convergence